



# Classification via local multi-resolution projections (extended version)

Jean-Baptiste Monnier

## ► To cite this version:

Jean-Baptiste Monnier. Classification via local multi-resolution projections (extended version). 2011.  
hal-00637010v2

**HAL Id: hal-00637010**

**<https://hal.science/hal-00637010v2>**

Submitted on 13 Dec 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Classification via local multi-resolution projections (extended version) Technical Report

Jean-Baptiste Monnier,  
e-mail: [monnier@math.jussieu.fr](mailto:monnier@math.jussieu.fr)  
Université Paris Diderot, Paris 7,  
LPMA, office 5B01,  
175 rue du Chevaleret,  
75013, Paris, France.

December 13, 2011

## Abstract

We focus on the supervised binary classification problem, which consists in guessing the label  $Y$  associated to a co-variate  $X \in \mathbb{R}^d$ , given a set of  $n$  independent and identically distributed co-variables and associated labels  $(X_i, Y_i)$ . We assume that the law of the random vector  $(X, Y)$  is unknown and the marginal law of  $X$  admits a density supported on a set  $\mathcal{A}$ . In the particular case of plug-in classifiers, solving the classification problem boils down to the estimation of the regression function  $\eta(X) = \mathbb{E}[Y|X]$ . Assuming first  $\mathcal{A}$  to be known, we show how it is possible to construct an estimator of  $\eta$  by localized projections onto a multi-resolution analysis (MRA). In a second step, we show how this estimation procedure generalizes to the case where  $\mathcal{A}$  is unknown. Interestingly, this novel estimation procedure presents similar theoretical performances as the celebrated local-polynomial estimator (LPE). In addition, it benefits from the lattice structure of the underlying MRA and thus outperforms the LPE from a computational standpoint, which turns out to be a crucial feature in many practical applications. Finally, we prove that the associated plug-in classifier can reach super-fast rates under a margin assumption.

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 62G05, 62G08; Secondary 62H30, 62H12.

KEY-WORDS: Nonparametric regression; Random design; Multi-resolution analysis; Supervised binary classification; Margin assumption.

# 1 Introduction

## 1.1 Setting

The supervised binary classification problem is directly related to a wide range of applications such as spam detection or assisted medical diagnosis (see [28, chap. 1] for more details). It can be described as follows.

*The supervised binary classification problem.* Let  $\mathcal{E}$  stand for a subset of  $\mathbb{R}^d$  and write  $\mathcal{Y} = \{0, 1\}$ . Assume we observe  $n$  co-variables  $X_i \in \mathcal{E}$  and associated labels  $Y_i \in \mathcal{Y}$  such that the elements of  $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$  are  $n$  independent realizations of the random vector  $(X, Y) \in \mathcal{E} \times \mathcal{Y}$  of unknown law  $\mathbb{P}_{X,Y}$ . Given  $\mathcal{D}_n$  and a new co-variate  $X_{n+1}$ , we want to predict the associated label  $Y_{n+1}$  so as to minimize the probability of making a mistake.

In other words, we want to build a **classifier**  $h_n : \mathcal{E} \mapsto \mathcal{Y}$  upon the data  $\mathcal{D}_n$ , which minimizes  $\mathbb{P}(h_n(X) \neq Y | \mathcal{D}_n)$ . It is well known that the Bayes classifier  $h^*(\tau) := \mathbf{1}_{\{\eta(\tau) \geq 1/2\}}$ , where  $\eta(\tau) := \mathbb{E}[Y | X = \tau] = \mathbb{P}(Y = 1 | X = \tau)$  (unknown in practice), is optimal among all classifiers since, for any other classifier  $h_n$ , we have  $\ell(h_n, h^*) := \mathbb{P}(h_n(X) \neq Y | \mathcal{D}_n) - \mathbb{P}(h^*(X) \neq Y) \geq 0$  (see [15]). As a consequence, we measure the classification risk  $\mathcal{T}(h_n)$  associated to a classifier  $h_n$  as its average relative performance over all data sets  $\mathcal{D}_n$ ,  $\mathcal{T}(h_n) = \mathbb{E}^{\otimes n} \ell(h_n, h^*)$ . As described in [15, Chap. 7], there is no classifier  $h_n$  such that  $\mathcal{T}(h_n)$  goes to zero with  $n$  at a specified rate for all distributions  $\mathbb{P}_{X,Y}$ . We therefore make the assumption that  $\mathbb{P}_{X,Y}$  belongs to a class of distributions  $\mathcal{P}$  (as large as possible) and aim at constructing a classifier  $h_n$  such that

$$\inf_{\theta_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{T}(\theta_n) \lesssim \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{T}(h_n) \lesssim (\log n)^\delta \inf_{\theta_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{T}(\theta_n), \quad n \geq 1, \quad (1)$$

where the infimum is taken over all measurable maps  $\theta_n$  from  $\mathcal{E}$  into  $\mathcal{Y}$  and  $\lesssim$  means lesser or equal up to a multiplicative constant factor independent of  $n$ . Any classifier  $h_n$  verifying eq. (1) will be said to be **(nearly) minimax optimal** when  $\delta = 0$  ( $\delta > 0$ ).  $\mathcal{P}$  will stand for the set of all distributions such that the marginal law  $\mathbb{P}_X$  of  $X$  admits a density  $\mu$  on  $\mathcal{E}$  and  $\eta$  belongs to a given smoothness class. Throughout the paper, we will denote by  $\mu$  the density of  $\mathbb{P}_X$ .

Many classifiers have been suggested in the literature, such as  $k$ -nearest neighbors, neural networks, support vector machine (SVM) or decision trees (see [15, 28]). In this paper, we will exclusively focus on **plug-in classifiers**  $h_n(\tau) := \mathbf{1}_{\{\eta_n(\tau) \geq 1/2\}}$ , where  $\eta_n$  stands for an estimator of  $\eta$ . With such classifiers, it is shown in [52] that,

$$\mathcal{T}(h_n) \leq 2\mathbb{E}^{\otimes n} \mathbb{E} |\eta_n(X) - \eta(X)|, \quad (2)$$

where the term on the rhs is known as the regression loss (of the estimator  $\eta_n$  of  $\eta$ ) in  $\mathbb{L}_1(\mathcal{E}, \mu)$ -norm. Eq. (2) shows in particular that rates of convergence on the classification risk of a plug-in classifier  $h_n$  can be readily derived from rates of convergence on the regression loss of  $\eta_n$ . This prompts us to focus on the regression problem, which can be stated in full generality as follows.

*The regression on a random design problem.* Let  $\mathcal{E}, \mathcal{Y}$  stand for subsets of  $\mathbb{R}^d$  and  $\mathbb{R}$ , respectively. Assume we dispose of  $n$  co-variables  $X_i \in \mathcal{E}$  and associated observations  $Y_i \in \mathcal{Y}$  such that the elements of  $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$  are  $n$  independent realizations of the random vector  $(X, Y) \in \mathcal{E} \times \mathcal{Y}$  of unknown law  $\mathbb{P}_{X,Y}$ . We define  $\xi := Y - \eta(X)$ , where  $\eta(\tau) := \mathbb{E}[Y|X = \tau]$ , so that by construction  $\mathbb{E}[\xi|X] = 0$ . Given  $\mathcal{D}_n$  and under the assumption that  $\mathbb{P}_{X,Y}$  belongs to a large class of distributions  $\mathcal{P}$ , we want to come up with an estimator  $\eta_n$  of  $\eta$ , which is as accurate as possible for the wide range of losses  $\mathcal{S}_p(\eta_n) = \mathbb{E}^{\otimes n} \mathbb{E} |\eta_n(X) - \eta(X)|^p$ ,  $p \geq 1$ .

As described previously, in the particular case where  $\mathcal{Y} = \{0, 1\}$ , we fall back on the regression problem associated to the classification problem with plug-in classifiers. In this case,  $\xi$  is bounded such that  $|\xi| \leq 1$ . Notice however that the regression on a random design problem stated above permits for  $\mathcal{Y}$  to be any subset of  $\mathbb{R}$  (including  $\mathbb{R}$  itself). To be more precise, and by analogy with eq. (1), our aim is to build an estimator  $\eta_n$  of  $\eta$  such that, for all  $p \geq 1$ ,

$$\inf_{\theta_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{S}_p(\theta_n) \lesssim \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{S}_p(\eta_n) \lesssim (\log n)^\delta \inf_{\theta_n} \sup_{\mathbb{P}_{X,Y} \in \mathcal{P}} \mathcal{S}_p(\theta_n), \quad n \geq 1, \quad (3)$$

where the infimum is taken over all measurable maps  $\theta_n$  from  $\mathcal{E}$  into  $\mathcal{Y}$ . And  $\eta_n$  will be said to be (nearly) minimax optimal when  $\delta = 0$  ( $\delta > 0$ ).

## 1.2 Motivations

Many estimators  $\eta_n$  of  $\eta$  have been suggested in the literature to solve the regression on a random design problem. Among them, the celebrated local polynomial estimator (LPE) has been praised for its flexibility and strong theoretical performances (see [48, 49]). As is well known, the LPE is minimax optimal in any dimension  $d \in \mathbb{N}$  and for any  $\mathcal{S}_p$ -loss,  $p \in (0, \infty]$ , over the set of laws  $\mathcal{P}$  such that (i)  $\mu$  is bounded from above and below on its support  $\mathcal{A} := \text{Supp} \mu = \{\tau : \mu(\tau) > 0\}$ , (ii)  $\eta$  belongs to a Hölder ball  $\mathcal{C}^s(\mathcal{E}, M)$  of radius  $M$  and (iii)  $\xi$  has sub-Gaussian tails. As a drawback, the LPE is computationally expansive since it requires to perform a new regression at every single point  $x \in \mathcal{A}$  where we want to estimate  $\eta$ .

Computational efficiency is however of primary importance in many practical applications. In this paper, we show that it is possible to construct a novel estimator  $\eta_n$  of  $\eta$  by localized projections onto multi-resolution analysis (MRA) of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  (where  $\lambda$  stands for the Lebesgue measure on  $\mathcal{E}$ ), which presents similar theoretical performances and is computationally more efficient than the LPE.

## 1.3 The hypotheses

In this section, we summarize the assumptions on  $\mu$ ,  $\mathcal{A}$ ,  $\eta$  and  $\xi$  that will be used throughout the paper.

*Assumption on  $\mu$ .* Let us denote by  $\mu_{\min}, \mu_{\max}$  two real numbers such that  $0 < \mu_{\min} \leq \mu_{\max} < \infty$ . As is standard in the regression on a random design setting, we assume that the density  $\mu$  is bounded above and below on its support  $\mathcal{A}$ .

(D1)  $\mu_{\min} \leq \mu(\tau) \leq \mu_{\max}$  for all  $\tau \in \mathcal{A}$ .

This guarantees that we have enough information at each point  $x \in \mathcal{A}$  in order to estimate  $\eta$  with best accuracy. For a study with weaker assumptions on  $\mu$ , the reader is referred to [20, 22], for example, and the references therein.

*Assumption on  $\mathcal{A}$ .* We first assume that,

(S1)  $\mathcal{A} = \mathcal{E} = [0, 1]^d$ .

Therefore  $\mathcal{A}$  is known under (S1). We will deal with the case where  $\mathcal{A}$  is unknown in Section 9.

*Assumption on  $\eta$ .* Fix  $r \in \mathbb{N}$ . In the sequel, we will assume that,

(H<sub>s</sub><sup>r</sup>) The regression function  $\eta$  belongs to the generalized Lipschitz ball  $\mathcal{L}^s(\mathcal{E}, M)$  of radius  $M$ , for some  $s \in (0, r)$ .

Unless otherwise stated,  $s$  is **unknown** but belongs to the interval  $(0, r)$ , where  $r$  is **known**. For a detailed review of generalized Lipschitz classes, the reader is referred to Section 13.1 below.

*Assumptions on the noise  $\xi$ .* We will consider the two following assumptions,

(N1) Conditionally on  $X$ , the noise  $\xi$  is uniformly bounded, meaning that there exists an absolute constant  $K > 0$  such that  $|\xi| \leq K$ .

(N2) The noise  $\xi$  is independent of  $X$  and normally distributed with mean zero and variance  $\sigma^2$ , which we will denote by  $\xi \sim \Phi(0, \sigma^2)$ .

Assumption (N1) is adapted to the supervised binary classification setting, where  $\mathcal{Y} = \{0, 1\}$ , while (N2) is more common in the regression on a random design setting, where  $\mathcal{Y} = \mathbb{R}$ .

*Combination of assumptions.* In the sequel, we will conveniently refer by (CS1) to the set of assumptions (D1), (S1), (N1) or (N2). As detailed below in Section 3, configuration (CS1) is comparable to what is customary in the regression on a random design setting.

## 2 Our results

Assuming at first  $\mathcal{A}$  to be known, we introduce a novel nonparametric estimator  $\eta^{\textcircled{a}}$  of  $\eta$  built upon local regressions against a multi-resolution analysis (MRA) of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  and show that, under (CS1), it is adaptive nearly minimax optimal over a wide generalized Lipschitz scale and across the wide range of losses  $\mathbb{L}_p(\mathcal{E}, \mu)$ ,  $p \in [1, \infty)$ . We subsequently show that these results generalize to the case where  $\mathcal{A}$  is unknown but belongs to a large class of (eventually disconnected) subsets of  $\mathbb{R}^d$ , provided we modify the estimator  $\eta^{\textcircled{a}}$  accordingly. We denote by  $\eta^{\textcircled{x}}$  this latter estimator and prove that  $\eta^{\textcircled{x}}$  can be used to build an adaptive nearly minimax optimal plug-in classifier, which can reach super-fast rates under a margin assumption. The above results essentially hinge on an exponential upper-bound on the probability of deviation of  $\eta^{\textcircled{a}}$  from  $\eta$  at a point, as detailed in Theorem 7.1. These results either improve on the current literature or are interesting in their own right for the following reasons.

- 1) They show that it is possible to use MRAs to construct an adaptive nearly minimax optimal estimator  $\eta^\circledast$  of  $\eta$  under the sole set of assumptions **(CS1)**. More precisely, our results (i) hold in any dimension  $d$ ; (ii) over the wide range of  $\mathbb{L}_p(\mathcal{E}, \mu)$ -losses,  $p \in [1, \infty)$ ; (iii) and a large Lipschitz scale; (iv) and do not require any assumption on  $\mu$  beyond **(D1)**. It is noteworthy that, in contrary to most alternative MRA-based estimation methods, no smoothness assumption on  $\mu$  is needed.
- 2) From a computational perspective,  $\eta^\circledast$  outperforms other estimators of  $\eta$  under **(D1)** since it takes full advantage of the lattice structure of the underlying MRA. In particular it requires at most as many regressions as there are data points to be computed everywhere on  $\mathcal{E}$ , while alternative kernel estimators must be recomputed at each single point of  $\mathcal{E}$ . We illustrate this latter feature through simulation.
- 3) Furthermore, and in contrary to alternative MRA-based estimators, the local nature of  $\eta^\circledast$  allows to relax the assumption that  $\mathcal{A}$  is known. This latter configuration allows for  $\mu$  to cancel on  $\mathcal{E}$  as long as it remains bounded on its support  $\mathcal{A}$ , which is particularly appropriate to the supervised binary classification problem under a margin assumption.
- 4) In the regression on a random design setting,  $\eta^\circledast$  bridges in fact the gap between usual linear wavelet estimators and alternative kernel estimators, such as the LPE. On the one hand,  $\eta^\circledast$  inherits its computational efficiency from the lattice structure of the underlying MRA. On the other hand, it features similar theoretical performances as the LPE in the random design setting. In particular, it remains a (locally) linear estimator of the data (modulo a spectral thresholding of the local regression matrix), and cannot discriminate finer smoothness than the one described by (generalized) Lipschitz spaces.

Here is the paper layout. We start by a literature review in [Section 3](#). We give a hand-waving introduction to the main ideas that underpin the local multi-resolution estimation procedure in [Section 4](#). We define notations that will be used throughout the paper and introduce MRAs in [Section 5](#). Our actual estimation procedure is described in [Section 6](#) and the results are detailed in [Section 7](#). We show how these results can be fine-tuned under additional assumptions in [Section 8](#). Assumption **(S1)** is relaxed and the properties of  $\eta^\star$  are detailed in [Section 9](#). We show how these latter results spread to the classification setting in [Section 10](#). Results of a simulation study with  $\eta^\circledast$  under **(CS1)** are given in [Section 11](#). Proofs of the regression results can be found in [Section 12](#). The proofs of the classification results are simple modifications of the proofs given in [\[4\]](#) and thus detailed in the Appendix in [Section 13](#). In addition, the Appendix contains a detailed review of generalized Lipschitz spaces and MRAs in [Section 13.1](#) and [Section 13.2](#), respectively.

### 3 Literature review

Both the regression on a random design problem and the classification problem have a long-standing history in nonparametric statistics. We will therefore limit ourselves to a brief account of the corresponding literature that is relevant to the present paper.

### 3.1 Classification with plug-in classifiers

Let us start with a review of some of the classification literature dedicated to plug-in classifiers. The seminal work [40] showed that plug-in rules are asymptotically optimal. It has been subsequently pointed out in [39] that the classification problem is in fact only sensitive to the behavior of  $\mathbb{P}_{X,Y}$  near the boundary line  $\mathcal{M} := \{\tau \in \mathcal{E} : \eta(\tau) = 1/2\}$ . So that assumptions on the behavior of  $\mathbb{P}_{X,Y}$  away from this boundary are in fact unnecessary. Subsequent works such as [3] have shown that convex combinations of plug-in classifiers can reach fast rates (meaning faster than  $n^{-1/2}$ , and thus faster than nonparametric estimation rates). More recently, it has been shown in [4] that plug-in classifiers can reach super fast rates (that is faster than  $n^{-1}$ ) under suitable conditions. All these results are derived under some sort of smoothness assumption on the regression function  $\eta$  (see [54]) and a margin assumption (**MA**) (see Section 10 for details). This latter assumption clarifies the behavior of  $\mathbb{P}_{X,Y}$  in a neighborhood of  $\mathcal{M}$  and kicks in naturally through the computation

$$\mathcal{J}(h_n) \leq \delta \mathbb{P}(0 < |2\eta(X) - 1| \leq \delta) + \mathbb{E}|\eta_n(X) - \eta(X)| \mathbf{1}_{\{|\eta_n(X) - \eta(X)| > \delta\}},$$

where  $\delta$  is chosen such that it balances the two terms on the rhs. Finally, [4] exhibited optimal convergence rates under smoothness and margin assumptions and showed that they are attained with plug-in classifiers. Let us now turn to the regression on a random design problem.

### 3.2 Regression on a random design with wavelets

First results on multi-resolution analysis (MRA) and wavelet bases (see [37, 41]) emerged in the nonparametric statistics literature in the early 1990's (see [30, 17, 16, 18, 19]). It has been proved that, under (**CS1**) and in the particular case where  $\mu$  is the uniform distribution on  $\mathcal{E}$ , thresholded wavelet estimators of  $\eta$  are nearly minimax optimal over a wide Besov scale and range of  $\mathbb{L}_p(\mathcal{E}, \mu)$ -losses (see [13]). In order to leverage on the power of MRAs and associated wavelet bases, several authors attempted to transpose these latter results to more general design densities  $\mu$ . This, however, led to a considerable amount of difficulties.

The literature relative to the study of wavelet estimators on an **unknown** random design breaks down into two main streams. (i) The first one aims at constructing new wavelet bases adapted to the (empirical) measure of the design (see [32, 33, 12, 50]). (ii) The second one aims at coming up with new algorithms to estimate the coefficients of the expansion of  $\eta$  on traditional wavelet bases (see [2, 26, 34, 44, 47]). The present paper belongs to this second line of research. As described in [26], the success of the LPE on a random design results from the fact that it is built as a “ratio”, which cancels out most of the influence of the design. In a wavelet context, a first suggestion has therefore been to use the ratio estimator of  $\eta$  (see [1, 45], for example), well known from the statistics literature on orthogonal series decomposition (see [23, 24] and [15, Chap. 17] and the references therein). Roughly speaking, the ratio estimator is the wavelet equivalent of the Nadaraya-Watson estimator (see [43, 53]). It is elaborated on the simple observation that  $\eta(x) = \eta(x)\mu(x)/\mu(x)$  for all  $x \in \mathcal{A}$ , where both  $g(\cdot) = \eta(\cdot)\mu(\cdot)$  and  $\mu(\cdot)$  are easily estimated via traditional wavelet methods. The ratio estimator relies thus unfortunately on the estimation of  $\mu$  itself and must therefore assume as much smoothness on  $\mu$  as on  $\eta$ .

To address that issue, an other approach has been introduced in [8, 31]. They work with  $d = 1$  and take  $\mathcal{E}$  to be the unit interval  $[0, 1]$ . Their approach relies on the wavelet estimation of  $\eta \circ G^{-1}$ , where  $G$  stands for the cumulative distribution of the design and  $G^{-1}$  for its generalized inverse. Results are therefore stated in term of regularity of  $f \circ G^{-1}$ . Unfortunately, this method does not readily generalize to the multi-dimensional case, where  $G$  admits no inverse.

Finally, [6] obtains adaptive near-minimax optimal wavelet estimators over a wide Besov scale under **(CS1)** by means of model selection techniques. His results are hence valid for the  $\mathbb{L}_2(\mathcal{E}, \mu)$ -loss only.

Other relevant references that proceed with hybrid estimators (LPE and kernel estimator or LPE and wavelet estimator) are [21] and [55]. They both work under **(CS1)**, with  $d = 1$  and assume that  $\mu$  is at least continuous.

## 4 A primer on local multi-resolution estimation under **(CS1)**

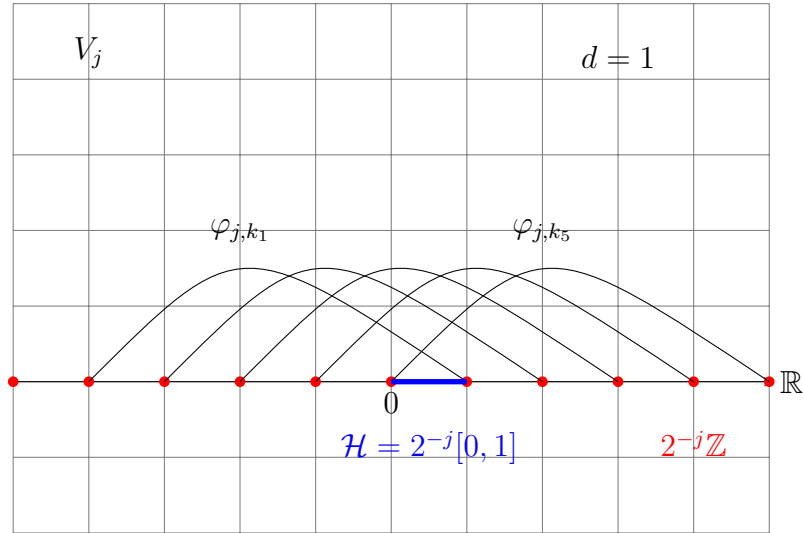


Figure 1: Description of the localization cells  $\mathcal{H}$  and their relations to the  $\text{Supp} \varphi_{j,k}$ .

In order to fix the ideas, let us now give a hand-waving introduction to the local multi-resolution estimation method. Throughout the paper, we will work with  $r$ -MRAs of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ , for some  $r \in \mathbb{N}$ , consisting of nested approximation spaces  $\mathcal{V}_j \subset \mathcal{V}_{j+1}$  built upon compactly supported scaling functions (see Section 5.2 and [11, 38, 10, 27]). Under the assumption that  $\eta$  belongs to the generalized Lipschitz ball  $\mathcal{L}^s(\mathcal{E}, M)$  of radius  $M$ , the essential supremum of the remainder of the orthogonal projection  $\mathcal{P}_j \eta$  of  $\eta$  onto  $\mathcal{V}_j$  decreases like  $2^{-js}$  (see Section 13.1). The regression function  $\eta$  can therefore be legitimately approximated by  $\mathcal{P}_j \eta$ . As an element of  $\mathcal{V}_j$ ,  $\mathcal{P}_j \eta$  may be written as an infinite linear combination of scaling functions at level  $j$ . In particular, there



exists a partition  $\mathcal{F}_j$  of  $\mathcal{E}$  into hypercubes of edge-length  $2^{-j}$  such that, for all  $\mathcal{H} \in \mathcal{F}_j$  and all  $x \in \mathcal{H}$ , we can write  $\mathcal{P}_j \eta(x) = \sum_{k \in \mathcal{S}_j(\mathcal{H})} \alpha_{j,k} \varphi_{j,k}(x)$ , where  $\mathcal{S}_j(\mathcal{H})$  stands for a finite subset of  $\mathbb{Z}^d$  (see Figure 1). This leaves us in turn with the estimation of coefficients  $(\alpha_{j,k})_{k \in \mathcal{S}_j(\mathcal{H})}$  for all  $\mathcal{H} \in \mathcal{F}_j$ , which is achieved by least-squares and provides us with the estimator  $\eta_j^\oplus$  of  $\eta$  on  $\mathcal{H}$ . It is noteworthy that the local estimator  $\eta_j^\oplus$  of  $\eta$  is exclusively built upon scaling functions and does not require the estimation of wavelet coefficients. In particular, it does not involve any sort of wavelet coefficient thresholding. To the best of the author knowledge, this is the first time that this local estimation procedure is proposed and studied from both a theoretical and computational perspective. In addition, we show that Lepski's method (see [35], for example) can be used to adaptively choose the resolution level  $j$ . Notice that Lepski's method has already been used in a MRA setting in [46]. In what follows, we detail the local multi-resolution estimation method and establish the near minimax optimality of  $\eta^\oplus$ .

## 5 Notations

### 5.1 Preliminary notations

In the sequel, we will denote by  $\mathcal{B}_p(z, \rho)$  the closed  $\ell_p$ -ball of  $\mathbb{R}^d$  of center  $z$  and radius  $\rho$ . More generally, we adopt the following notations: for any subset  $\mathcal{S}$  of a topological space  $\mathcal{E}$ ,  $\text{Closure}(\mathcal{S})$  will stand for its closure,  $\text{Inter}(\mathcal{S})$  for its interior,  $\mathcal{S}^c$  for its complement in  $\mathcal{E}$  and  $\partial\mathcal{S} = \text{Closure}(\mathcal{S}) \cap \text{Inter}(\mathcal{S})^c$  for its boundary. Besides, for any two  $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}^d$ , we will denote the relative complement of  $\mathcal{S}_2$  in  $\mathcal{S}_1$  by  $\mathcal{S}_1 \setminus \mathcal{S}_2 (= \mathcal{S}_1 \cap \mathcal{S}_2^c)$  and their symmetric difference  $\mathcal{S}_1 \setminus \mathcal{S}_2 \cup \mathcal{S}_2 \setminus \mathcal{S}_1$  by  $\mathcal{S}_1 \Delta \mathcal{S}_2$ . For any subset  $\mathcal{S}$  of  $\mathbb{R}^d$ ,  $z \in \mathbb{R}^d$  and  $\tau \in \mathbb{R}^+$ , we will write  $z + \mathcal{S}$  and  $\tau\mathcal{S}$  to mean the sets  $\{z + u : u \in \mathcal{S}\}$  and  $\{\tau u : u \in \mathcal{S}\}$ , respectively. Finally, given a set (of functions)  $\mathcal{R}$ ,  $\text{Span}\mathcal{R}$  will denote the set of finite linear combinations of elements of  $\mathcal{R}$ .

For any  $p \in \mathbb{N}$ , vectors  $v$  of  $\mathbb{R}^p$  will be seen as elements of  $\mathcal{M}_{p,1}$ , that is matrix with  $p$  rows and one column. For any two  $u, v \in \mathbb{R}^p$ ,  $\langle u, v \rangle$  will denote their Euclidean scalar product. In addition, for any  $p, q \in \mathbb{N}$  and  $M \in \mathcal{M}_{p,q}$ ,  $M^t$  will stand for the transpose of  $M$ . For any two matrices  $M, P$ ,  $M \cdot P$  will denote their matrix product when it makes sense.  $[M]_{k,\ell}$  and  $[M]_{k,\bullet}$  will respectively stand for the element of  $M$  located at line  $k$ , column  $\ell$  and the  $k^{\text{th}}$  row of  $M$ . Finally,  $\|M\|_S$  will denote the spectral norm of  $M$  (see [29, §5.6.6]).

We denote by  $\lfloor z \rfloor$  the integer part of  $z \in \mathbb{R}$  defined as  $\max\{a \in \mathbb{Z} : a \leq z\}$ . More generally, given  $z \in \mathbb{R}^d$ , we write  $\lfloor z \rfloor$  the integer part of  $z$ , meant in a coordinate-wise sense. In the same way, we denote by  $\lceil z \rceil$  the smallest integer greater than  $z$  (in a coordinate-wise sense). We write rhs (resp. lhs) to mean *right-* (resp. *left-*) *hand-side* and sometimes write  $:=$  to mean *equal by definition*. Throughout the paper, we will refer to constants independent of  $n$  as *absolute constants* and  $c, C$  will stand for absolute constants whose value may vary from line to line. For any two sequences  $a_n, b_n$  of  $n$ , we will write  $a_n \lesssim b_n$  to mean  $a_n \leq Cb_n$  for some absolute constant  $C$  and  $a_n \approx b_n$  to mean that there exist two constants  $c, C$  independent of  $n$  such that  $cb_n \leq a_n \leq Cb_n$ .

## 5.2 The polynomial reproduction property

In what follows, we will exclusively consider MRAs built upon Daubechies' scaling functions  $\varphi_{j,k}$  (see [Section 13.2](#)). Given a natural integer  $r$ , we will refer by  $r$ -MRA to a MRA whose nested approximation spaces  $\mathcal{V}_j$  reproduce polynomials up to order  $r - 1$ . Daubechies' scaling functions are appealing in the estimation framework since they are compactly supported and have minimal volume supports among scaling functions that give rise to  $r$ -MRAs. Recall that a  $r$ -MRA can explain Lipschitz smoothness  $s$  for any  $s \in (0, r)$ . Obviously, the usefulness of the 1-MRA, also known as the Haar MRA, is fairly limited when dealing with the estimation of smooth functions. However, the simple geometry of the Haar scaling function brings a great deal of simplification in the proofs, which will allow us to highlight the key points of our estimation procedure and hopefully clarify our arguments. This is the unique reason why we have treated the two cases  $r = 1$  and  $r \geq 2$  distinctly in the proofs. The reader is referred to the Appendix in [Section 13.2](#) for a detailed review of MRAs of  $\mathbb{R}^d$ .

## 5.3 General notations

Consider the Daubechies'  $r$ -MRA of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  built upon Daubechies' scaling function  $\varphi$ , as described in [Section 13.2](#). We will denote by  $\text{Supp}\varphi_{j,k} = \{\tau \in \mathbb{R}^d : \varphi_{j,k}(\tau) > 0\}$  the support of  $\varphi_{j,k}$ . Recall that  $\text{Supp}\varphi = [-(r-1), r]^d$ . To alleviate notations, we will write  $\varphi_k$  in place of  $\varphi_{0,k}$  and  $\varphi_j$  in place of  $\varphi_{j,0}$ . Notice that  $\text{Closure}(\text{Supp}\varphi_{j,k})$  is in fact a closed hyper-cube of  $\mathbb{R}^d$  whose corners lie on the lattice  $2^{-j}\mathbb{Z}^d$  (see [Section 13.2](#)). Moreover, for any  $x \in \mathcal{A}$ , write

$$\mathcal{S}_j(x) = \{\nu \in \mathbb{Z}^d : x \in \text{Supp}\varphi_{j,\nu}\}.$$

Furthermore, we write  $\mathcal{F}_j := 2^{-j}((0, 1)^d + \mathbb{Z}^d) \cap \mathcal{E}$ . It defines a partition of  $\mathcal{E}$  into  $2^{jd}$  hypercubes of edge length  $2^{-j}$ , modulo a  $\lambda$ -null set. For the sake of concision, we write  $R = 2r - 1$  in the sequel. We have the following proposition, whose proof is straightforward and thus left to the reader.

**Proposition 5.1.**  *$\mathcal{S}_j$  verifies the following properties,*

1.  *$\mathcal{S}_j$  is constant on each element  $\mathcal{H} \in \mathcal{F}_j$ . We will denote by  $\mathcal{S}_j(\mathcal{H})$  its value on  $\mathcal{H}$ .*
2. *Moreover, for any two  $\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{F}_j$ ,  $\mathcal{H}_1 \neq \mathcal{H}_2$ ,  $\mathcal{S}_j(\mathcal{H}_1)$  differs from  $\mathcal{S}_j(\mathcal{H}_2)$  by at least one element.*
3. *Finally, for any  $\mathcal{H} \in \mathcal{F}_j$ ,  $\#\mathcal{S}_j(\mathcal{H}) = R^d$*

It is a direct consequence of [Proposition 5.1](#) that in the case where  $r = 1$ , we have  $\#\mathcal{S}_j(\mathcal{H}) = 1$  for all  $\mathcal{H} \in \mathcal{F}_j$ . We denote its single element by  $\nu(\mathcal{H})$ . It is in fact easy to show that  $\nu(\mathcal{H}) = \lfloor 2^j x \rfloor$  for any  $x \in \mathcal{H}$ . For any  $\mathcal{H} \in \mathcal{F}_j$ , we write

$$\begin{aligned} \alpha_{\mathcal{H}} &= (\alpha_{j,\nu})_{\nu \in \mathcal{S}_j(\mathcal{H})} \in \mathbb{R}^{R^d}, \\ \varphi_{\mathcal{H}}(\cdot) &= (\varphi_{j,\nu}(\cdot) \mathbf{1}_{\mathcal{H}}(\cdot))_{\nu \in \mathcal{S}_j(\mathcal{H})} \in \mathbb{R}^{R^d}. \end{aligned}$$

and denote by  $Y_{\mathcal{H}} = (Y_i \mathbf{1}_{\mathcal{H}}(X_i))_{1 \leq i \leq n}$ . Moreover, we write  $\mathcal{I}_{\mathcal{H}} = \{i : X_i \in \mathcal{H}\}$  and  $n_{\mathcal{H}} = \#\mathcal{I}_{\mathcal{H}}$ .

## 6 Construction of the local estimator $\eta^\circledast$

Assume we are under **(CS1)** and work with the Daubechies'  $r$ -MRA of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  (see [Section 13.2](#)). The estimation procedure is local, so that we start by selecting a point  $x \in \mathcal{A}$  (see [Section 13.3](#) for a technical remark). By construction, there exists  $\mathcal{H} \in \mathcal{F}_j$  such that  $x \in \mathcal{H}$ . We want to estimate  $\eta$  at point  $x$ . As detailed in [Section 13.2](#), an estimator of  $\eta$  can be reduced to an estimator of the orthogonal projection  $\mathcal{P}_j\eta$  of  $\eta$  onto  $\mathcal{V}_j$ , modulo an error  $\mathcal{R}_j\eta$ , such that  $|\mathcal{R}_j\eta| \leq M2^{-js}$  when  $\eta$  belongs to the generalized Lipschitz ball  $\mathcal{L}^s(\mathcal{E}, M)$  of radius  $M$ . Now, we can write

$$\mathcal{P}_j\eta(x) = \sum_{k \in \mathbb{Z}^d} \alpha_{j,k} \varphi_{j,k}(x) = \sum_{k \in \mathcal{S}_j(\mathcal{H})} \alpha_{j,k} \varphi_{j,k}(x) = \langle \alpha_{\mathcal{H}}, \varphi_{\mathcal{H}}(x) \rangle.$$

This leaves us with exactly  $R^d$  coefficients  $\alpha_{j,\nu}, \nu \in \mathcal{S}_j(\mathcal{H})$  to estimate, which are valid for any  $x \in \mathcal{H}$ . We evaluate these coefficients by least-squares. Denote by  $B_{\mathcal{H}} \in \mathcal{M}_{n,R^d}$  the matrix whose rows are the vectors  $\varphi_{\mathcal{H}}(X_i)^t$  for  $1 \leq i \leq n$ . Let us denote by  $k_1, \dots, k_{R^d}$  the elements of  $\mathcal{S}_j(\mathcal{H})$ . Then we choose

$$\begin{aligned} \alpha_{\mathcal{H}}^\diamond &\in \arg \min_{a \in \mathbb{R}^{R^d}} \sum_{i=1}^n \left( Y_i - \sum_{t=1}^{R^d} a_t \varphi_{j,k_t}(X_i) \right)^2 \mathbb{1}_{\mathcal{H}}(X_i) \\ &= \arg \min_{a \in \mathbb{R}^{R^d}} \|Y_{\mathcal{H}} - B_{\mathcal{H}} \cdot a\|_{\ell_2(\mathbb{R}^n)}^2, \end{aligned} \quad (4)$$

where we set  $\alpha_{\mathcal{H}}^\diamond = 0$  if the arg min above contains more than one element. Let us write  $Q_{\mathcal{H}} = B_{\mathcal{H}}^t \cdot B_{\mathcal{H}} / n \in \mathcal{M}_{R^d, R^d}$ . As is well known, when  $Q_{\mathcal{H}}$  is invertible, the arg min on the rhs of [eq. \(4\)](#) admits one single element which writes as follows,

$$\alpha_{\mathcal{H}}^\diamond = Q_{\mathcal{H}}^{-1} \cdot \frac{1}{n} B_{\mathcal{H}}^t \cdot Y_{\mathcal{H}}. \quad (5)$$

Naturally, we will denote the corresponding estimator of  $\mathcal{P}_j\eta$  at point  $x$  by  $\eta_{\mathcal{H}}^\diamond(x) = \langle \alpha_{\mathcal{H}}^\diamond, \varphi_{\mathcal{H}}(x) \rangle$ . We now introduce a thresholded version of  $\eta_{\mathcal{H}}^\diamond$  based on the spectral thresholding of  $Q_{\mathcal{H}}$ . We denote by  $\lambda_{\min}(Q_{\mathcal{H}})$  the smallest eigenvalue of  $Q_{\mathcal{H}}$  in the case where  $r \geq 2$ , when  $Q_{\mathcal{H}}$  is actually a matrix, and  $Q_{\mathcal{H}}$  itself in the case where  $r = 1$ , when it is a real number. Furthermore, we define

$$\eta_{\mathcal{H}}^\circledast(x) = \begin{cases} 0 & \text{if } \pi_n^{-1} > \lambda_{\min}(Q_{\mathcal{H}}) \\ \eta_{\mathcal{H}}^\diamond(x) & \text{otherwise} \end{cases}, \quad (6)$$

where  $\pi_n$  is a tuning parameter. In practice, and unless otherwise stated, we choose  $\pi_n = \log n$ . Moreover, we assume throughout the paper that  $n$  is large enough so that  $\pi_n^{-1} \leq \min(\frac{g_{\min}}{2}, 1)$ , where, for reasons that will be clarified later, we have denoted,

$$g_{\min} := \mu_{\min} c_{\min}, \quad (7)$$

and  $c_{\min}$  stands for the strictly positive constant defined in the proof of [Proposition 12.4](#). Ultimately, the estimator  $\eta_j^{\textcircled{a}}$  of  $\mathcal{P}_j\eta$  is defined as,

$$\eta_j^{\textcircled{a}}(x) = \sum_{\mathcal{H} \in \mathcal{F}_j} \eta_{\mathcal{H}}^{\textcircled{a}}(x) \mathbf{1}_{\mathcal{H}}(x), \quad x \in \mathcal{E}. \quad (8)$$

## 7 The results

Let  $r$  be a natural integer, denote by  $\mathcal{P}$  the set of all distributions on  $\mathcal{E} \times \mathcal{Y}$  and write

$$\mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r) := \{\mathbb{P} \in \mathcal{P} : (\mathbf{CS1}) \text{ and } (\mathbf{H}_s^r) \text{ hold true}\}. \quad (9)$$

Furthermore, we define  $j_r, j_s, J$  and  $t(n)$  such that,

$$\begin{aligned} 2^{j_r} &= \lfloor n^{\frac{1}{2r+d}} \rfloor, & 2^{j_s} &= \lfloor n^{\frac{1}{2s+d}} \rfloor, \\ 2^{J_d} &= \lfloor nt(n)^{-2} \rfloor, & t(n)^2 &= \kappa \pi_n^2 \log n, \end{aligned}$$

where  $\kappa$  is a positive real number to be chosen later. In addition, we write  $\mathcal{J}_n = \{j_r, j_r + 1, \dots, J-1, J\}$ . Notice that  $j_s$  strikes the balance between bias and variance in the sense that, for  $\log n \geq (2s+d) \log 2$  and  $s \in (0, r)$ , one has got

$$n^{-\frac{1}{2}} 2^{j_s \frac{d}{2}} \leq 2^{-j_s s}, \quad (10a)$$

$$2^{-j_s s} \leq 2^{r+\frac{d}{2}} 2^{j_s \frac{d}{2}} n^{-\frac{1}{2}}, \quad (10b)$$

$$2^{-j_s s} \leq 2^r n^{-\frac{s}{2s+d}}. \quad (10c)$$

Throughout the sequel, we assume that  $n$  is large enough so that the latter inequalities hold true. Our first result gives an upper bound on the probability of deviation of  $\eta_j^{\textcircled{a}}$  from  $\eta$  at a point  $x \in \mathcal{A}$ .

**Theorem 7.1.** *Fix  $r \in \mathbb{N}$  and assume we are under  $(\mathbf{CS1})$  and  $(\mathbf{H}_s^r)$ . Recall that  $\eta_j^{\textcircled{a}}$  is defined in [eq. \(8\)](#). Then, for all  $j \in \mathcal{J}_n$ , all  $\delta > 2M2^{-j_s} \max(1, 3\pi_n R^d \mu_{\max})$  and all  $x \in \mathcal{A}$ , we have got*

$$\begin{aligned} & \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r)} \mathbb{P}^{\otimes n}(|\eta(x) - \eta_j^{\textcircled{a}}(x)| \geq \delta) \\ & \leq 2R^{2d} \exp\left(-n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max} R^{4d} + \frac{4}{3} R^{2d} \pi_n^{-1}}\right) \mathbf{1}_{\{\delta \leq M\}} + R^d \Lambda\left(\frac{\delta 2^{-j \frac{d}{2}}}{2\pi_n R^d}\right), \end{aligned} \quad (11)$$

where  $\Lambda$  is defined as follows,

$$\Lambda(\delta) = \begin{cases} 2 \exp\left(-\frac{n\delta^2}{18K^2\mu_{\max} + 4K2^{j\frac{d}{2}}\delta}\right), & \text{under (N1)} \\ 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right) \right\} \\ + 2 \exp\left(-\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right), & \text{under (N2)} \end{cases}$$

As a consequence of the above theorem, we can deduce the (near) minimax optimality of  $\eta_{j_s}^{\otimes}$  over generalized Lipschitz balls.

**Corollary 7.1.** *Fix  $r \in \mathbb{N}$  and assume we are under (CS1) and  $(\mathbf{H}_s^r)$ . Then, for any  $p \in [1, \infty)$  and  $j \in \mathcal{J}_n$ , one has got*

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r)} \mathbb{E}^{\otimes n} \|\eta - \eta_j^{\otimes}\|_{\mathbb{L}_p(\mathcal{E}, \mu)}^p \leq C(p) \pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p, \quad (12)$$

where  $\eta_j^{\otimes}$  and  $C(p)$  are defined in [eq. \(8\)](#) and [Proposition 12.1](#) below, respectively. A fortiori, when  $s$  is **known**, we can choose  $j = j_s$  and apply [eq. \(10a\)](#) and [eq. \(10c\)](#) above to obtain

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r)} \mathbb{E}^{\otimes n} \|\eta - \eta_{j_s}^{\otimes}\|_{\mathbb{L}_p(\mathcal{E}, \mu)}^p \leq C(p) 2^{rp} \pi_n^p n^{-\frac{sp}{2s+d}}.$$

This, together with the lower-bound of [Theorem 7.3](#), proves that  $\eta_{j_s}^{\otimes}$  is (nearly) minimax optimal over the generalized Lipschitz ball  $\mathcal{L}^s(\mathcal{E}, M)$  of radius  $M$ .

The next Theorem shows that the approximation level  $j$  can be determined from the data  $\mathcal{D}_n$  so that we obtain adaptation over a wide generalized Lipschitz scale.

**Theorem 7.2.** *Fix  $r \in \mathbb{N}$  and assume we are under (CS1) and  $(\mathbf{H}_s^r)$ . We define*

$$g(j, k) := \left( \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n) + \frac{2^{k\frac{d}{2}}}{\sqrt{n}} t(n) \right),$$

$$j^{\otimes}(x) := \inf\{j \in \mathcal{J}_n : |\eta_j^{\otimes}(x) - \eta_k^{\otimes}(x)| \leq g(j, k), \forall k \in \mathcal{J}_n, k > j\}, \quad x \in \mathcal{A},$$

where  $\eta_j^{\otimes}$  is defined in [eq. \(8\)](#) and  $\inf \emptyset = \max(\mathcal{J}_n) = J$ . If  $\kappa$  is chosen large enough, meaning  $\kappa \geq \frac{p}{2} C_9^{-1}$ , where  $C_9$  is defined in [Proposition 12.2](#), then we obtain

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r)} \mathbb{E}^{\otimes n} \|\eta_{j^{\otimes}(\cdot)}^{\otimes}(\cdot) - \eta(\cdot)\|_{\mathbb{L}_p(\mathcal{E}, \mu)}^p \leq 5^p 2^{rp} t(n)^p n^{-\frac{sp}{2s+d}}.$$

So that  $\eta_{j^{\otimes}(\cdot)}^{\otimes}(\cdot)$  is a nearly minimax adaptive estimator of  $\eta$  over the generalized Lipschitz scale

$$\bigcup_{0 < s < r} \mathcal{L}^s(\mathcal{E}, M).$$

Finally, we prove that  $\eta^\circledast$  is indeed (nearly) minimax optimal by giving the corresponding lower-bound result.

**Theorem 7.3.** *Assume we are under (CS1) and ( $\mathbf{H}_s^r$ ). We write  $\inf_{\theta_n}$  the infimum over all estimators  $\theta_n$  of  $\eta$ , that is all measurable functions of the data  $\mathcal{D}_n$ . Then, for  $d \geq 1$ ,  $s > 0$ , we have, for all  $1 \leq p < \infty$ ,*

$$\inf_{\theta_n} \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS1}, \mathbf{H}_s^r)} \mathbb{E}^{\otimes n} \|\theta_n - \eta\|_{\mathbb{L}_p(\mathcal{E}, \mu)}^p \gtrsim n^{-\frac{sp}{2s+d}}.$$

The next section shows how these results can be improved in the case where we benefit from additional information on  $\mu$  or  $\eta$ .

## 8 Refinement of the results

As can be seen from [Corollary 7.1](#) and [Theorem 7.2](#) above,  $\pi_n$  appears as a multiplicative factor in the upper-bounds and thus deteriorates them by a multiplicative  $\log n$  term. However, this needs not be the case, and under appropriate additional assumptions,  $\pi_n$  can be chosen to be a constant. Consider indeed the following two assumptions.

(O1) We know  $\mu_{\min}^* \in \mathbb{R}$ , such that  $0 < \mu_{\min}^* \leq \mu_{\min}$ .

(O2) We know a finite positive real number  $M$  such that  $\|\eta\|_{\mathbb{L}_\infty(\mathcal{E}, \lambda)} \leq M$ .

Under (O1), we know a lower bound  $\mu_{\min}^*$  of  $\mu_{\min}$ , and therefore a lower bound  $g_{\min}^*$  of  $g_{\min}$  (see [eq. \(7\)](#)). Under (O1), we will thus choose  $\pi_n^{-1} = \min(\frac{g_{\min}^*}{2}, 1)$ . It is straightforward to show that [Theorem 7.1](#) is still valid with this new value of  $\pi_n$  (see [Remark 12.1](#) in the proof of [Theorem 7.1](#)), and thus all the subsequent results follow as well. Under (O2), we know an upper bound  $M$  of the essential supremum of  $\eta$  on  $\mathcal{E}$ . In that case, we redefine

$$\eta_{\mathcal{H}}^\circledast(x) = T_M(\eta_{\mathcal{H}}^\diamond(x)) \mathbb{1}_{\{\lambda_{\min}(Q_{\mathcal{H}}) > 0\}}, \quad (13)$$

where, for any  $z \in \mathbb{R}$ , we have written  $T_M(z) = z \mathbb{1}_{\{|z| \leq M\}} + M \text{sign}(z) \mathbb{1}_{\{|z| > M\}}$ . Once again, it is straightforward to show that [Theorem 7.1](#) is now valid with  $\pi_n^{-1} = \min(\frac{g_{\min}^*}{2}, 1)$  and  $2M$  in place of  $M$  in the indicator function on the rhs of [eq. \(11\)](#) (see [Remark 12.1](#) in the proof of [Theorem 7.1](#)), and thus all the subsequent results follow as well.

Notice that  $\pi_n$  is an absolute constant under (O1) and (O2), while it is an increasing sequence of  $n$  to be fine-tuned by the statistician otherwise. Hence  $\pi_n$  appears to be the price to pay for not knowing a lower bound of  $\mu_{\min}$  or an upper bound of the essential supremum of  $\eta$  on  $\mathcal{E}$ .

## 9 Relaxation of assumption (S1)

### 9.1 The problem

Now, we would like to relax assumption (S1) and allow for  $\mathcal{A}$  to be an unknown subset of  $\mathcal{E}$ , eventually disconnected. Under (CS1), the success of  $\eta^\circledast$  stems from the fact that it is

constructed upon an approximation grid of the form  $2^{-j}\mathbb{Z}^d \cap [0,1]^d$ , whose edges coincide exactly with the boundary of  $\mathcal{A}$ . In the case where  $\mathcal{A}$  is unknown, some cells of the lattice might straddle the boundary of  $\mathcal{A}$  and thus require a new treatment.

In order to handle this new configuration, we will need to make a smoothness assumption on the boundary of  $\mathcal{A}$  and allow for the estimation cells to move with the point at which we want to estimate  $\eta$ . Ultimately, we devise a new estimator  $\eta^{\star}$  of  $\eta$  which is built upon a moving approximation grid. In fact, this new estimation method ensures that the point  $x$  at which we want to estimate  $\eta$  always belongs to a cell  $\mathcal{H}$  of  $\mathcal{F}_j$  at resolution level  $j$ , whose center belongs to  $\mathcal{A}$ . This will ensure that local regressions performed on cells that straddle the boundary of  $\mathcal{A}$  are still meaningful.

The smoothness assumption we will make on  $\mathcal{A}$  might be compared to the support assumption made in [4, eq. (2.1)] in the classification context. In substance, it is assumed in [4] that  $\mathcal{A}$  is locally ball-shaped to be compatible with the ball-shaped support of the LPE kernel, which they use to estimate  $\eta$ . In our case, we perform estimation with multi-dimensional scaling functions whose supports are cube-shaped and will thus assume that  $\mathcal{A}$  is locally cube-shaped.

## 9.2 Smoothness assumption on $\mathcal{A}$

Let us now make these informal arguments more precise. To that end we introduce assumption (S2) as an alternative to (S1) above. Fix an absolute constant  $\mathbf{m}_0 \in (0,1)$  and recall that  $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$ . With these notations, (S2) goes as follows,

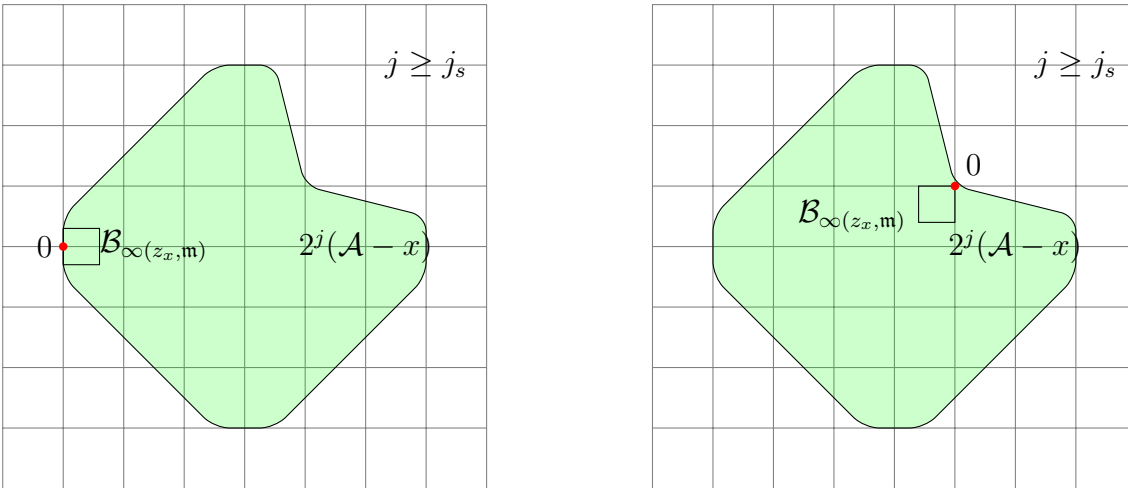


Figure 2: (S2) allows for  $\mathcal{A}$  to be non-convex and eventually disconnected.

(S2)  $\mathcal{E} = \mathbb{R}^d$  and  $\mathcal{A}$  belongs to  $\mathcal{A}_{j_s}$ , where

$$\begin{aligned} \mathcal{A}_{j_s} := \{ \mathcal{A} \subset \mathbb{R}^d : \exists \mathbf{m} \geq \mathbf{m}_0, \forall x \in \mathcal{A}, \\ \exists z_x \in \mathbb{R}^d, 0 \in \mathcal{B}_{\infty}(z_x, \mathbf{m}) \subset 2^{j_s}(\mathcal{A} - x) \}, \end{aligned}$$

In words, **(S2)** means that if we zoom close enough to any  $x \in \mathcal{A}$ , we can find a hypercube  $\mathcal{B}_\infty(z_x, \mathbf{m})$  that contains  $x$  and is a subset of  $\mathcal{A}$ . Notice readily that for all  $j_1 \geq j_2$ , the component of  $2^{j_2}(\mathcal{A} - x)$  that contains 0 is a subset of the component of  $2^{j_1}(\mathcal{A} - x)$  that contains 0, so that  $\mathcal{A}_{j_2} \subset \mathcal{A}_{j_1}$ . Therefore  $\mathcal{A}_{j_s}$  grows with  $n$  and shrinks with  $s$ . Of course, **(S1)** is a particular case of **(S2)**. Setting **(S2)** allows  $\mathcal{A}$  to be unknown and belong to a wide class of subsets of  $\mathbb{R}^d$ , eventually disconnected (see Figure 2).

In the sequel, we will conveniently refer by **(CS2)** to the set of assumptions **(D1)**, **(S2)**, **(N1)** or **(N2)**.

### 9.3 Moving local estimation under (CS2)

As detailed above,  $\eta^{\star}$  is obtained by local regression on a moving approximation grid. Let us describe the construction of  $\eta^{\star}$  more precisely.

First of all, we split the sample into two pieces. For simplicity, let us assume that we dispose of  $2n$  data points. The first half of the sample points, which we denote by  $\mathcal{D}'_n = \{(X'_i, Y'_i), i = 1, \dots, n\}$ , will be used to identify the support  $\mathcal{A}$  of  $\mu$ , while the second half, which we denote by  $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$ , will be used to estimate the scaling functions coefficients by local regressions.

Let us denote by  $\mathcal{H}_0$  the cell  $2^{-j}[0, 1]^d$  of the lattice  $2^{-j}\mathbb{Z}^d$  at resolution  $j$ . And denote by  $\mathcal{H}_0(x)$  the same cell centered in  $x$ , that is  $\mathcal{H}_0(x) = x - 2^{-j-1} + 2^{-j}[0, 1]^d$ . Then, the construction of  $\eta^{\star}_j(x)$  at a point  $x \in \mathbb{R}^d$  goes as follows. (i) If none of the design points  $(X'_i)$  of the sample  $\mathcal{D}'_n$  lie in  $\mathcal{H}_0(x)$ , then take  $\eta^{\star}(x) = 0$ . (ii) If one or more design points of the sample  $\mathcal{D}'_n$  lie in  $\mathcal{H}_0(x)$ , we select one of them and denote it by  $X'_{i_x}$  (the selection procedure is of no importance beyond computational considerations). By construction,  $x$  belongs to the cell  $\mathcal{H}_0(X'_{i_x})$  centered in  $X'_{i_x} \in \mathcal{A}$ . Since  $X'_{i_x}$  belongs to  $\mathcal{A}$ , it makes sense to perform a local regression on  $\mathcal{H}_0(X'_{i_x})$  with the sample points  $\mathcal{D}_n$ , which gives rise to an estimator  $\eta^{\star}$  of  $\eta$  valid at any point of  $\mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$ . It is noteworthy that this procedure uses the sample  $\mathcal{D}'_n$  to identify the support  $\mathcal{A}$  of  $\mu$ .

Interestingly, the above estimation procedure requires at most as many regressions as there are data points in  $\mathcal{D}'_n$  to return an estimator  $\eta^{\star}$  of  $\eta$  at every single point  $x \in \mathcal{A}$ . It is therefore computationally more efficient than any other kernel estimator, such as the LPE. The computational performance of  $\eta^{\star}$  can in fact be further improved in the sense that the local regression on the cell  $\mathcal{H}_0(X'_{i_x})$  can be omitted if the cell  $\mathcal{H}_0(X'_i)$  is itself included in the union of cells centered at other design points of  $\mathcal{D}'_n$ . In particular, we can choose  $X'_{i_x}$  to be a design point  $X'_i$  of  $\mathcal{D}'_n$  that belongs to  $\mathcal{H}_0(x)$  and for which a local regression has already been performed, if it exists, or any one of the  $X'_i$  that belong to  $\mathcal{H}_0(x)$  otherwise.

Intuitively, the computational efficiency of  $\eta^{\star}$  stems from the fact that the design points  $(X'_i)$  provide some valuable information on the unknown support  $\mathcal{A}$  of  $\mu$ , which can be exploited under **(CS2)**. In particular, and as we will see below, **(D1)** guarantees that the design points of  $\mathcal{D}'_n$  populate  $\mathcal{A}$  densely enough so that, as long as  $j \leq J$ , the cells  $\mathcal{H}_0(X'_i)$ ,  $1 \leq i \leq n$ , form a cover of  $\mathcal{A}$ , modulo a set whose  $\mu$ -measure decreases almost exponentially fast toward zero with  $n$ .



## 9.4 Construction of the local estimator $\eta^{\star}$

Assume we are under **(S2)** and work with the Daubechies'  $r$ -MRA of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ . Obviously, shifting the approximation grid is equivalent to shifting the data points  $(X_i)$  of  $\mathcal{D}_n$  and keeping the lattice fixed. For ease of notations and clarity, we adopt this second point of view. In order to compute  $\eta^{\star}$  at a point  $x \in \mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$ , we want to shift the design points in such a way that  $X'_{i_x}$  falls right in the middle of  $\mathcal{H}_0$ . In other words, we want  $X'_{i_x}$  to be shifted at point  $2^{-j-1} \in \mathbb{R}^d$  (whose coordinates are worth  $2^{-j-1} \in \mathbb{R}$ ). This corresponds to the change of variable  $\tilde{X}_i = X_i - (X'_{i_x} - 2^{-j-1})$ , where we have denoted by  $X_i$  and  $\tilde{X}_i$  the representations of a same data point in the canonical and shifted coordinate systems of  $\mathbb{R}^d$ , respectively. In order to compute  $\eta^{\star}$  at point  $x \in \mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$ , it is therefore enough to perform a local regression on  $\mathcal{H}_0$  against the shifted data points,

$$\tilde{\mathcal{D}}_x = \{(\tilde{X}_i, Y_i), i = 1, \dots, n\}.$$

For the sake of concision, we will denote by  $\tilde{u} = u - (X'_{i_x} - 2^{-j-1})$  the coordinate representation of a point  $u$  in the shifted coordinate system of  $\mathbb{R}^d$ . Let us denote by  $k_1, \dots, k_{R^d}$  the elements of  $\mathcal{S}_j(\mathcal{H}_0)$ . With these notations, eq. (4) must be corrected and written as

$$\alpha_{\mathcal{H}_0}^{\diamond} \in \arg \min_{a \in \mathbb{R}^{R^d}} \sum_{i=1}^n \left( Y_i - \sum_{t=1}^{R^d} a_t \varphi_{j, k_t}(\tilde{X}_i) \right)^2 \mathbb{1}_{\mathcal{H}}(\tilde{X}_i), \quad (14)$$

where we set  $\alpha_{\mathcal{H}_0}^{\diamond} = 0$  if the argmin above contains more than one element. The notations introduced in Section 5.3 can be updated to this new setting as follows.  $B_{\mathcal{H}_0}$  stands now for the random matrix of  $\mathcal{M}_{n, R^d}$  whose rows are the  $\varphi_{\mathcal{H}_0}(\tilde{X}_i)^t$ ,  $i = 1, \dots, n$ . In addition, we recall that we have defined  $Q_{\mathcal{H}_0} = B_{\mathcal{H}_0}^t \cdot B_{\mathcal{H}_0} / n \in \mathcal{M}_{R^d, R^d}$ . Its coefficients write thus as

$$[Q_{\mathcal{H}_0}]_{\nu, \nu'} = \frac{1}{n} \sum_{i=1}^n \varphi_{j, \nu}(\tilde{X}_i) \varphi_{j, \nu'}(\tilde{X}_i) \mathbb{1}_{\mathcal{H}_0}(\tilde{X}_i), \quad \nu, \nu' \in \mathcal{S}_j(\mathcal{H}_0).$$

Notice here that  $\mathcal{S}_j(\mathcal{H}_0) = \{\nu \in \mathbb{Z}^d : 2^{-1} \in \text{Supp} \varphi_{\nu}\}$ , which neither depends on  $j$  nor  $x$ . Therefore, and for later reference, we denote

$$\mathfrak{S} := \{\nu \in \mathbb{Z}^d : 2^{-1} \in \text{Supp} \varphi_{\nu}\}, \quad (15)$$

In addition, if we write  $Y_{\mathcal{H}_0} = (Y_i \mathbb{1}_{\mathcal{H}_0}(\tilde{X}_i))_{1 \leq i \leq n}$ , then eq. (5) still holds true when the solution to eq. (14) is unique. So that, for all  $x \in \mathcal{H}_0(X'_{i_x}) \cap \mathcal{A}$ , we can write  $\eta_{\mathcal{H}_0}^{\diamond}(\tilde{x}) = \langle \alpha_{\mathcal{H}_0}^{\diamond}, \varphi_{\mathcal{H}_0}(\tilde{x}) \rangle$ . Finally eq. (6) remains valid with  $X_i$  replaced by  $\tilde{X}_i$  and  $\mathcal{H}$  by  $\mathcal{H}_0$ ,  $\eta_{\mathcal{H}_0}^{\circledast}$  redefined as  $\eta_{\mathcal{H}_0}^{\star}$  and  $g_{\min}$  redefined as

$$g_{\min} = \mu_{\min} c_{\min}, \quad (16)$$

where  $c_{\min}$  is the strictly positive constant defined in Lemma 12.1 below. So that ultimately, the estimator  $\eta_j^{\star}$  of  $\mathcal{P}_j \eta$  at a point  $x \in \mathbb{R}^d$  writes as

$$\eta_j^{\star}(x) = \eta_{\mathcal{H}_0}^{\star}(\tilde{x}), \quad x \in \mathcal{E}. \quad (17)$$

Notice that by contrast with [eq. \(8\)](#) above, the sum over the hypercubes of  $\mathcal{F}_j$  has disappeared. This is due to the fact that the approximation grid moves with  $x$  so that we end up virtually always performing estimation on the same hypercube  $\mathcal{H}_0$ .

## 9.5 The results

Interestingly,  $\eta^{\star}$  still verifies similar results as the ones described in [Section 7](#). To be more precise, recall that we work with a sample of size  $2n$  broken up into two pieces  $\mathcal{D}_n$  and  $\mathcal{D}'_n$  of size  $n$ . Let us redefine  $\mathcal{J}_n$  so that  $\mathcal{J}_n = \{j_s, j_s + 1, \dots, J - 1, J\}$  where  $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$ . Then, we obtain the following result in place of [Theorem 7.1](#).

**Theorem 9.1.** *Fix  $r \in \mathbb{N}$  and assume we are under (CS2) and  $(\mathbf{H}_s^r)$ . Recall that  $\eta_j^{\star}$  is defined in [eq. \(17\)](#). Then, for all  $j \in \mathcal{J}_n$ , all  $\delta > 2M2^{-j_s} \max(1, 3\pi_n R^d \mu_{\max})$  and all  $x \in \mathcal{A}$ , we have got*

$$\begin{aligned} & \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r)} \mathbb{P}^{\otimes n}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta) \\ & \leq 3R^{2d} \exp\left(-n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}}\right) \mathbb{1}_{\{\delta \leq M\}} \\ & + R^d \Lambda\left(\frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right), \end{aligned}$$

where  $\Lambda$  has been defined in [Theorem 7.1](#).

Left aside the fact that  $\eta^{\star}$  is constructed upon a sample of size  $2n$ , the sole difference with the result of [Theorem 7.1](#) is that the leading constant in front of the exponential on the second line has changed from  $2R^d$  to  $3R^d$ . Furthermore, it is straightforward to deduce from [Theorem 9.1](#) results similar to [Corollary 7.1](#), [Theorem 7.2](#) and [Theorem 7.3](#), and a fortiori the refined results obtained in [Section 8](#), for  $\eta^{\star}$  under (CS2). The proofs of these results for  $\eta^{\star}$  under the set of assumptions (CS2) follow, for the most part, exactly the same lines as the proofs given for  $\eta^{\circledast}$  under (CS1). Details can be found in [Section 12.2](#).

## 10 Classification via local multi-resolution projections

Recall from [\[4\]](#) that the margin assumption can be written as,

(MA) There exist constants  $C_* > 0$  and  $\vartheta \geq 0$  such that

$$\mathbb{P}(0 < |2\eta(X) - 1| \leq t) \leq C_* t^{\vartheta}, \quad \forall t > 0.$$

The binary classification setting corresponds to (CS2), under assumptions (N1) and (O2). Notice besides that we have  $K = 1$  in (N1) and  $M = 1$  in  $(\mathbf{H}_s^r)$ . Since we are under (O2), it follows from [Section 8](#) that  $\pi_n = \pi_0 = \min(1, \frac{g_{\min}}{2})$  is independent of  $n$  and  $\eta^{\star}$  is capped at  $M = 1$  as in [eq. \(13\)](#). For the sake of coherence, we denote by  $j^{\star}$  the adaptive resolution level

built upon  $\eta^{\star}$ , as described in [Theorem 7.2](#), and define  $\mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r)$  by analogy with [eq. \(9\)](#) above. Finally, we recall that  $\eta^{\star}$  is built upon a sample of size  $2n$  split into two sub-samples  $\mathcal{D}_n$  and  $\mathcal{D}'_n$  of size  $n$ . In this context, we have the following Proposition.

**Proposition 10.1.** *Fix  $r \in \mathbb{N}$  and assume we are in the binary classification setting. Assume moreover that  $(\mathbf{H}_s^r)$  holds true. Then, for all  $j \in \mathcal{J}_n$ , all  $x \in \mathcal{A}$  and any  $1 \geq \delta > 2 \cdot 2^{-js} \max(1, 3\pi_0 R^d \mu_{\max})$ , we have got*

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r)} \mathbb{P}^{\otimes n}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta) \\ \leq 3R^{2d} n^{-\kappa C'_{10}} + 2R^d \exp(-C_{10} a_n(j) \delta^2), \end{aligned} \quad (18)$$

and for  $1 \geq \delta > 62^r t(n) n^{-\frac{s}{2s+d}}$ ,

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r)} \mathbb{P}^{\otimes n}(|\eta(x) - \eta_{j^{\star}(x)}^{\star}(x)| \geq \delta) \\ \leq 11R^{2d} (\log n)^2 n^{-\kappa C'_6} + 2R^d \exp(-C_{11} n^{\frac{2s}{2s+d}} \delta^2), \end{aligned} \quad (19)$$

where we have written  $a_n(j) := n2^{-jd}$  and the constants  $C'_6, C'_{10}, C_{10}, C_{11}$  are all defined in the proof, [eq. \(38\)](#) and [eq. \(39\)](#).

The proof of this latter result is a straightforward consequence of [Theorem 7.1](#) and [Theorem 7.2](#). It is detailed in the Appendix. All the results that follow can be easily derived from [\[4\]](#). Their proofs are therefore not reported in this Section but delayed to the Appendix as well. The first result derives an upper-bound on the classification loss from an upper bound on the probability of deviation of a classifier  $\eta_n$  from  $\eta$  at a point  $x \in \mathcal{A}$ . It is similar to [\[4, Theorem 3.1\]](#) in scope and proof.

**Theorem 10.1.** *Fix  $r \in \mathbb{N}$  and assume we are in the binary classification setting. Assume moreover that  $(\mathbf{H}_s^r)$  and  $(\mathbf{MA})$  hold true. Let  $\eta_n$  be an estimator of  $\eta$  and assume there exists  $\delta_n \in (0, 1)$  and strictly positive constants  $C_{11}, C_{12}, C_{13}$  such that, for all  $1 \geq \delta \geq \delta_n$  and almost all  $x \in \mathcal{A}$ , we have*

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r, \mathbf{MA})} \mathbb{P}^{\otimes n}(|\eta_n(x) - \eta(x)| \geq \delta) \leq C_{11} b_n + C_{12} \exp(-C_{13} a_n \delta^2), \quad (20)$$

for some positive sequences  $a_n$  and  $b_n$ . Consider the plug-in classifier  $h_n(\cdot) := \mathbb{1}_{\{\eta_n(\cdot) \geq \frac{1}{2}\}}$ . Then, we have got

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r, \mathbf{MA})} \mathcal{J}(h_n) &\leq 2C_* \delta_n^{1+\vartheta} \left( 1 + C_{11} b_n \delta_n^{-(1+\vartheta)} \right. \\ &\quad \left. + C_{12} \sum_{k \geq 1} 2^{k(1+\vartheta)} \exp(-C_{13} a_n \delta_n^2 2^{2(k-1)}) \right). \end{aligned}$$

As a consequence of [Theorem 10.1](#), we can use the plug-in classifier built upon  $\eta^{\star}$  to obtain similar results as the ones given in [\[4, Lemma 3.1\]](#) for LPE based plug-in classifiers.

**Corollary 10.1.** Fix  $r \in \mathbb{N}$  and assume we are in the binary classification setting. Assume moreover that  $(\mathbf{H}_s^r)$  and  $(\mathbf{MA})$  hold true. Consider the plug-in classifiers  $h_{j_s}^\star(\cdot) = \mathbb{1}_{\{\eta_{j_s}^\star(\cdot) \geq \frac{1}{2}\}}$  and  $h_{j_s^\star}^\star(\cdot) = \mathbb{1}_{\{\eta_{j_s^\star}^\star(\cdot) \geq \frac{1}{2}\}}$ . Then, we have

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r, \mathbf{MA})} \mathcal{T}(h_{j_s}^\star) \leq C_{15} n^{-\frac{s}{2s+d}(1+\vartheta)}, \quad \text{as soon as } \kappa > \frac{1+\vartheta}{2C'_{10}}, \quad (21)$$

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r, \mathbf{MA})} \mathcal{T}(h_{j_s^\star}^\star) \leq C_{16} (\log n)^{\frac{1+\vartheta}{2}} n^{-\frac{s}{2s+d}(1+\vartheta)}, \quad \text{as soon as } \kappa > \frac{1+\vartheta}{2C'_6}, \quad (22)$$

where  $C'_6, C'_{10}$  have been defined in Proposition 10.1 above and  $C_{15}, C_{16}$  are defined in the proof in eq. (40) and eq. (42).

In addition, the lower bound of [4, Theorem 3.5] carries over to the present setting after minor modifications. We have indeed the following Theorem.

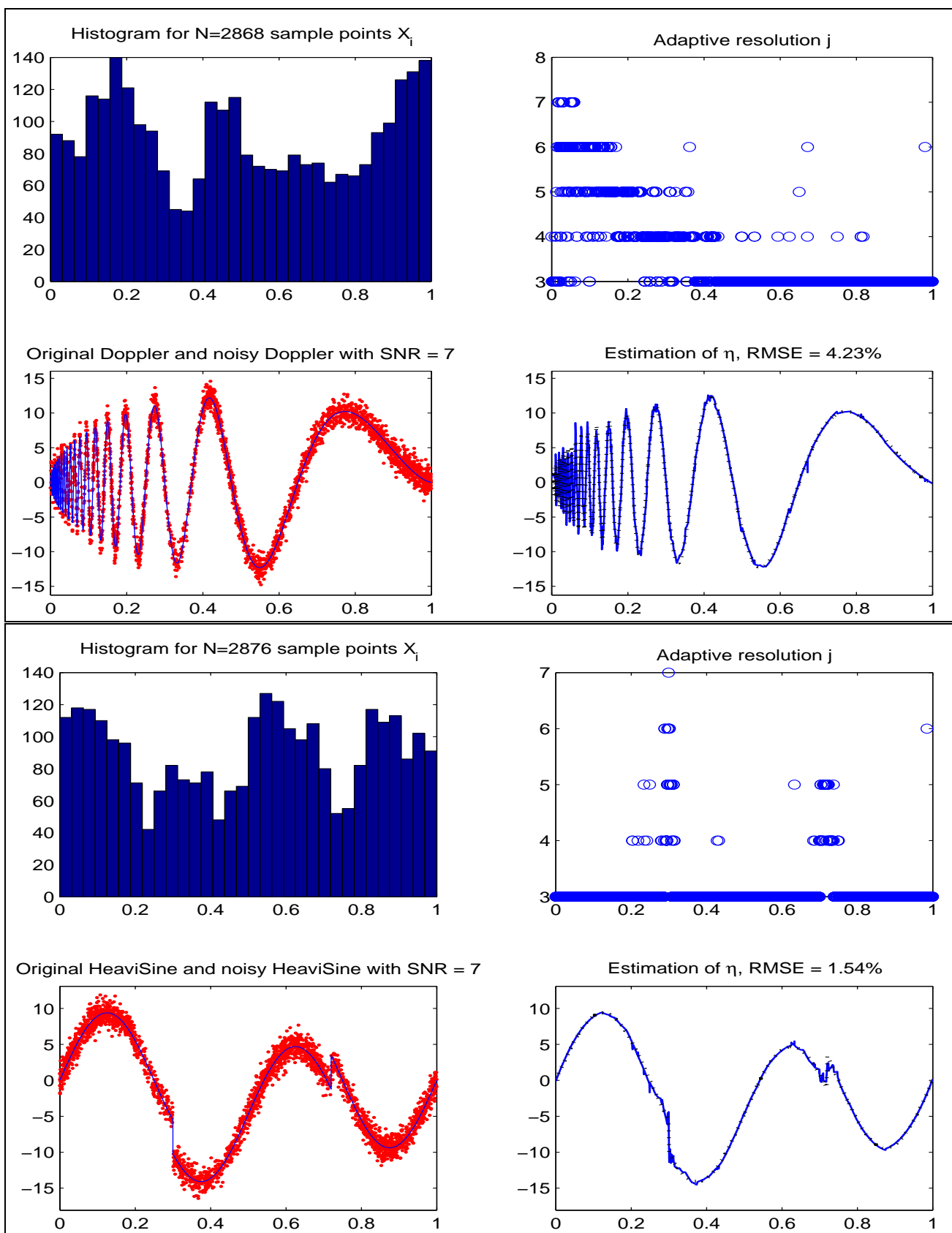
**Theorem 10.2.** Fix  $r \in \mathbb{N}$  and assume we are in the binary classification setting. Assume moreover that  $(\mathbf{H}_s^r)$  and  $(\mathbf{MA})$  hold true and  $s\vartheta \leq d$ . Then there exists a constant  $c > 0$  such that for any classifier  $\theta_n$ , that is any  $\{0, 1\}$ -valued measurable function  $\theta_n$  of the data  $\mathcal{D}_n$ , we have

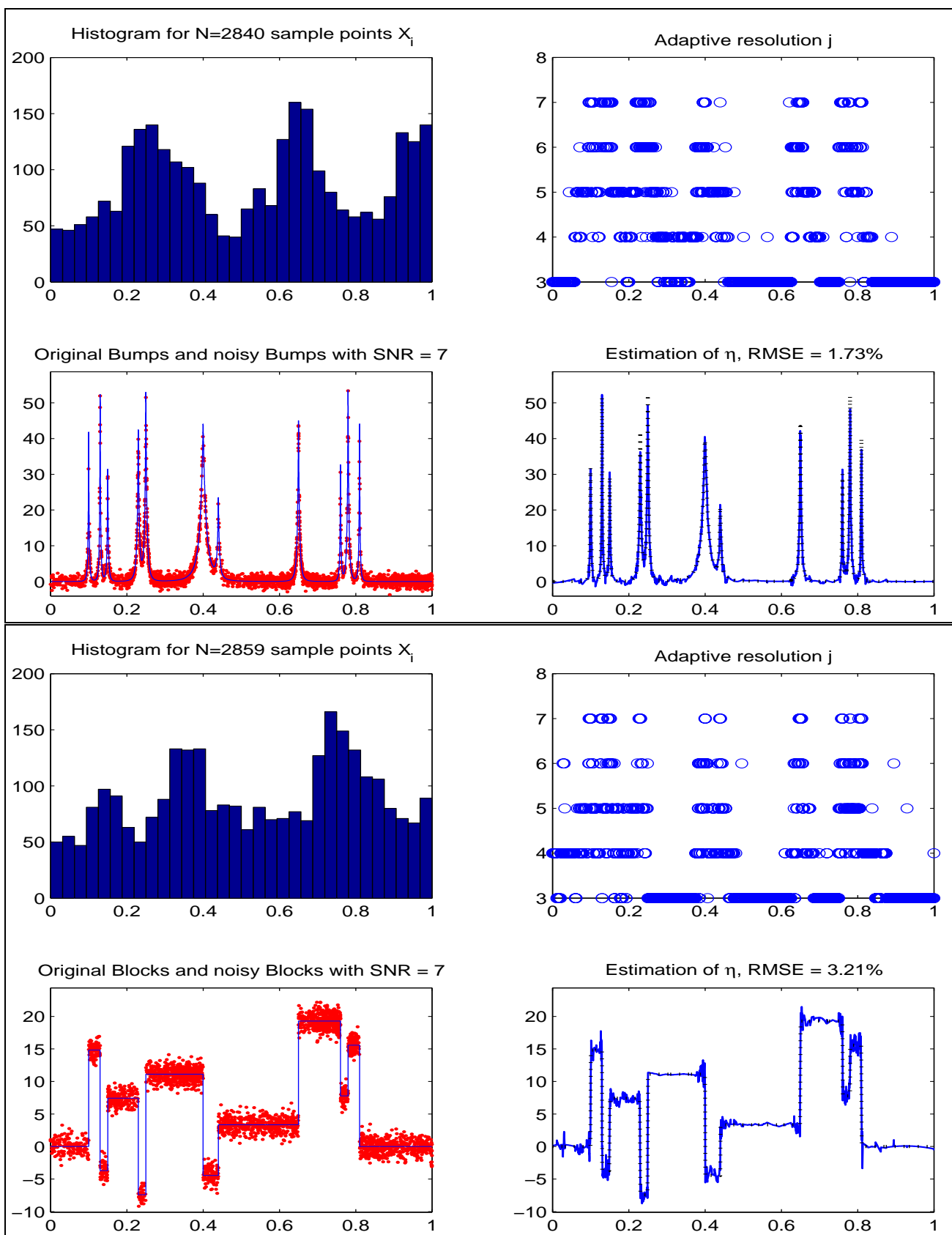
$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r, \mathbf{MA})} \mathcal{T}(\theta_n) \gtrsim n^{-\frac{s}{2s+d}(1+\vartheta)}.$$

So that, under the above assumptions, the classifiers  $h^\star$  defined in Corollary 10.1 are (nearly) minimax optimal.

## 11 Simulation study

In order to illustrate the performance of  $\eta_{j_s}^\circ$ , we have carried out a simulation study in the regression setting in the one-dimensional case, that is with  $d = 1$ . As detailed earlier, the sole purpose of this simulation is to show that (1)  $\eta^\circ$  can be easily implemented and is computationally efficient, (2)  $\eta^\circ$  works well in practice in the case where the density of the design  $\mu$  is discontinuous, (3) and to give an intuitive visual feel for  $\eta^\circ$ , which is built upon the juxtaposition of local regressions against a set of scaling functions. In particular, we run our simulation against benchmark signals, which allows to compare them with the ones detailed in the literature for alternative kernel estimators (see simulation study in [35], for example). We have run them under  $(\mathbf{CS1})$ , which corresponds to the case where  $\eta_j^\circ$  can be completely computed with exactly  $2^j$  regressions. We have in particular  $\mathcal{E} = [0, 1] = \mathcal{A}$ . We focus on the functions  $\eta$  introduced in [17] and used as a benchmark in numerous subsequent simulation studies. They are made available through the Wavelab850 library freely available at <http://www-stat.stanford.edu/~wavelab/>. In addition we have chosen the noise  $\xi$  to be standard normal, that is we are working under  $(\mathbf{N2})$  with  $\sigma = 1$ . In all cases, we have chosen the signal-to-noise ratio (SNR) to be equal to 7. To be more specific, we are working on a dyadic grid  $G$  of  $[0, 1]$  of resolution  $2^{-15}$ . We compute the root-mean-squared-error (RMSE) of both





the signal and the noise on that grid and rescale the signal so that its RMSE be seven times bigger than the one of the noise.

Let us now give details about the simulation of the sample points and the computation of the estimator. We divide the unit-interval into ten sub-segments  $A_k := 10^{-1}[k, k + 1]$  for  $k = 0, \dots, 9$ . We define the density of  $X$  as follows.

$$\mu(x) = \sum_{k=0}^9 p_k \lambda(A_k)^{-1} \mathbb{1}_{A_k}(x).$$

We choose the  $p_k$ 's at random. To that end, we denote by  $(u_k)_{0 \leq k \leq 9}$  ten realizations of the uniform random variable on  $[.25, 1]$ , write  $v = u_0 + \dots + u_9$  and set  $p_k = u_k v^{-1}$ . Notice that this guarantees that  $\mu \geq \min_{0 \leq k \leq 9} 10p_k \geq \mu_{\min} = 0.25$  on  $[0, 1]$ . We then simulate 3000 sample points  $X_i$  according to  $\mu$ . Finally, we bring the points back on the grid  $G$  by assimilating them to their nearest grid node. Since the  $X_i$ 's are supposed to be drawn from a law that is absolutely continuous with respect to the Lebesgue measure on  $[0, 1]$ , we must keep only one data point per grid node. This reduces the number of data points from 3000 to the number that is reported on top of each of the histograms.

In order to compute the adaptive estimator at sample points  $X_i$ , we use the boundary-corrected scaling functions coded into Wavelab850 for  $r = 3$  and for which we must have  $j \geq 3$ . We set  $J = \lceil \log(n/\log n)/\log 2 \rceil$ . The elimination of redundant sample points on the grid removes on average 150 points so that we obtain  $J = 10$ . We therefore have  $\mathcal{J}_n = \{3, 4, \dots, 10\}$ . Notice interestingly that the computation of  $\eta_3^{\textcircled{a}}$  requires only 8 regressions and  $\eta_{10}^{\textcircled{a}}$  requires 1,024 of them. This is much smaller than for the LPE whose computation necessitates as many regressions as there are sample points at each resolution level. In practice, we compute the minimum eigenvalues of all regression matrices across partitions and resolution levels and choose  $\pi_n^{-1}$  to be the first decile of this set of values. When proving theoretical results, we have chosen  $\eta_j^{\textcircled{a}}$  to be zero on the small probability event where the minimum eigenvalue of the regression matrix is smaller than  $\pi_n^{-1}$ . In practice we can choose it to be an average value of the nearby cells in order to get an estimator that is overall more appealing to the eye. In our simulation, we in fact do not use that modification. Instead, we modify  $j^{\textcircled{a}}$  to be the highest  $j \in \{3, \dots, j^{\textcircled{a}}\}$  such that  $\eta_j^{\textcircled{a}}$  has been computed from a valid regression matrix, meaning a regression matrix whose smallest eigenvalue is greater than the threshold  $\pi_n^{-1}$ .

In practice, for a given signal, we generate  $\mu$  at random and compute  $\eta_{j^{\textcircled{a}}}^{\textcircled{a}}$  for 100 samples drawn from  $\mu$ . We quantify the performance  $\eta_{j^{\textcircled{a}}}^{\textcircled{a}}$  by its relative RMSE, meaning its RMSE computed at sample points  $X_i$  divided by the amplitude of the true signal, that is its maximal absolute value on the underlying dyadic grid. We display results for “Doppler”, “HeaviSine”, “Bumps” and “Blocks” corresponding to the median performance among the 100 trials. Each figure displays four graphs. Clockwise from the top left corner, they display in turn, an histogram of sample points  $X_i$ ; the adaptive level  $j^{\textcircled{a}}$  at sample points  $X_i$ ; the true signal (black dots) and the estimator  $\eta_{j^{\textcircled{a}}}^{\textcircled{a}}$  at sample points  $X_i$  (solid blue line) and its corresponding relative RMSE in the title; and finally the original signal (solid blue line) with its noisy version at sample points  $X_i$  (red dots).

## 12 Proofs

### 12.1 Proof of the upper-bound results under (CS1)

#### 12.1.1 Proof of [Corollary 7.1](#)

Consider the term

$$I = \int_{\mathcal{A}} \mathbb{E}[|\eta(x) - \eta_j^{\circledast}(x)|^p] \mu(x) dx.$$

Now, apply [Proposition 12.1](#) and notice that  $\int_{\mathcal{A}} \mu(x) dx = 1$  to show that  $I$  is upper-bounded by the term that appears on the rhs of [eq. \(12\)](#) stated in [Corollary 7.1](#). In particular, for all  $1 \leq p < \infty$ , we obtain  $I \leq C(p) \pi_n^p t(n)^{-p} \leq C(p) < \infty$ . This in turn proves that we can apply the Fubini-Tonelli theorem to get

$$I = \mathbb{E}[\|\eta - \eta_j^{\circledast}\|_{\mathbb{L}_p(\mathcal{E}, \mu)}^p],$$

and concludes the proof.  $\square$

#### 12.1.2 Proof of [Theorem 7.1](#)

**Local estimation under (CS1) in the case where  $r = 1$ .** Let  $x \in \mathcal{A}$  and  $j \in \mathcal{J}_n$ . There exists  $\mathcal{H} \in \mathcal{F}_j$  such that  $x \in \mathcal{H}$ . Recall that with Haar MRAs, we have  $\mathcal{S}_j(\mathcal{H}) = \{\nu(\mathcal{H})\}$  so that  $B_{\mathcal{H}} \in \mathcal{M}_{n,1}$  and for any  $i \in \mathcal{I}_{\mathcal{H}}$ ,  $[B_{\mathcal{H}}]_i = \varphi_{j,\nu(\mathcal{H})}(X_i) \mathbb{1}_{\mathcal{H}}(X_i) = \varphi_{j,\nu(\mathcal{H})}(X_i)$  since  $\mathcal{H} = \text{Supp} \varphi_{j,\nu(\mathcal{H})}$ . So that

$$Q_{\mathcal{H}} = \frac{1}{n} \sum_{i \in \mathcal{I}_{\mathcal{H}}} \varphi_{j,\nu(\mathcal{H})}(X_i)^2 = 2^{jd} \frac{n_{\mathcal{H}}}{n}.$$

On  $\{Q_{\mathcal{H}} \geq \pi_n^{-1}\}$ , we can divide by  $Q_{\mathcal{H}}$  and thus write

$$\alpha_{j,\nu(\mathcal{H})}^{\diamond} = \frac{\sum_{i=1}^n Y_i \varphi_{j,\nu(\mathcal{H})}(X_i)}{\sum_{i=1}^n \varphi_{j,\nu(\mathcal{H})}(X_i)^2} = 2^{-jd/2} \frac{\sum_{i=1}^n Y_i \mathbb{1}_{\{X_i \in \text{Supp} \varphi_{j,\nu(\mathcal{H})}\}}}{\sum_{i=1}^n \mathbb{1}_{\{X_i \in \text{Supp} \varphi_{j,\nu(\mathcal{H})}\}}} = 2^{-jd/2} \frac{1}{n_{\mathcal{H}}} \sum_{i \in \mathcal{I}_{\mathcal{H}}} Y_i.$$

On this event, the estimation error can hence be written as

$$\begin{aligned} |\mathcal{P}_j \eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| &= 2^{jd/2} |\alpha_{j,\nu(\mathcal{H})} - \alpha_{j,\nu(\mathcal{H})}^{\diamond}| \\ &= \frac{1}{n_{\mathcal{H}}} \left| \sum_{i \in \mathcal{I}_{\mathcal{H}}} \alpha_{j,\nu(\mathcal{H})} \varphi_{j,\nu(\mathcal{H})}(X_i) - \sum_{i \in \mathcal{I}_{\mathcal{H}}} Y_i \right| \\ &= \frac{1}{n_{\mathcal{H}}} \left| \sum_{i \in \mathcal{I}_{\mathcal{H}}} (\mathcal{P}_j \eta(X_i) - Y_i) \right| \\ &= \frac{1}{n_{\mathcal{H}}} \left| \sum_{i \in \mathcal{I}_{\mathcal{H}}} (\mathcal{R}_j \eta(X_i) + \xi_i) \right| \end{aligned}$$



$$\begin{aligned}
&= \frac{n}{n_{\mathcal{H}}} 2^{-jd/2} \frac{1}{n} \left| \sum_{i=1}^n (\varphi_{j,\nu(\mathcal{H})}(X_i) (\mathcal{R}_j \eta(X_i) + \xi_i)) \right| \\
&\leq 2^{j\frac{d}{2}} \pi_n \frac{1}{n} \left| \sum_{i=1}^n (\varphi_{j,\nu(\mathcal{H})}(X_i) (\mathcal{R}_j \eta(X_i) + \xi_i)) \right|,
\end{aligned}$$

since  $Q_{\mathcal{H}}^{-1} \leq \pi_n$ . Write

$$W_{j,\nu(\mathcal{H})} = \frac{1}{n} \left| \sum_{i=1}^n (\varphi_{j,\nu(\mathcal{H})}(X_i) (\mathcal{R}_j \eta(X_i) + \xi_i)) \right|.$$

By definition, we have  $\eta_j^{\textcircled{a}}(x) = \eta_{\mathcal{H}}^{\textcircled{a}}(x)$ , so that we have

$$\begin{aligned}
\mathbb{P}(|\eta(x) - \eta_j^{\textcircled{a}}(x)| \geq \delta) &= \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{\textcircled{a}}(x)| \geq \delta, Q_{\mathcal{H}} \geq \pi_n^{-1}) \\
&\quad + \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{\textcircled{a}}(x)| \geq \delta, Q_{\mathcal{H}} < \pi_n^{-1}).
\end{aligned} \tag{23}$$

By construction,  $\eta_{\mathcal{H}}^{\textcircled{a}}(x) = \eta_{\mathcal{H}}^{\diamond}(x)$  on the event  $\{Q_{\mathcal{H}} \geq \pi_n^{-1}\}$  and  $\eta_{\mathcal{H}}^{\textcircled{a}}(x) = 0$  on its complement. So that we obtain  $|\eta(x) - \eta_{\mathcal{H}}^{\textcircled{a}}(x)| = |\eta(x)| \leq M$  on the rhs of eq. (23). Notice in addition that  $M2^{-js} \geq |\mathcal{R}_j \eta(x)|$  under  $(\mathbf{H}_{\mathbf{s}}^{\mathbf{r}})$  (see Section 13.1). Finally, we obtain, for  $\frac{\delta}{2} > M2^{-js} \geq |\mathcal{R}_j \eta(x)|$ ,

$$\begin{aligned}
&\mathbb{P}(|\eta(x) - \eta_j^{\textcircled{a}}(x)| \geq \delta) \\
&\leq \mathbb{P}(|\mathcal{P}_j \eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| \geq \frac{\delta}{2}, Q_{\mathcal{H}} \geq \pi_n^{-1}) + \mathbb{P}(Q_{\mathcal{H}} < \pi_n^{-1}) \mathbf{1}_{\{\bar{M} \geq \delta\}} \\
&\leq \mathbb{P}(W_{j,\nu(\mathcal{H})} \geq 2^{-jd/2-1} \delta \pi_n^{-1}) + \mathbb{P}(Q_{\mathcal{H}} < \pi_n^{-1}) \mathbf{1}_{\{\bar{M} \geq \delta\}}.
\end{aligned}$$

where we have written  $\bar{M} = M$ . The term on the rhs is tackled thanks to Proposition 12.5. Regarding the lhs, a direct application of Proposition 12.6 allows to write, for all  $\delta > 2M2^{-js} \max(1, 3\pi_n \mu_{\max})$ ,

$$\mathbb{P}(W_{j,\nu(\mathcal{H})} \geq 2^{-jd/2-1} \delta \pi_n^{-1}) \leq \Lambda(2^{-j\frac{d}{2}-1} \delta \pi_n^{-1}).$$

This concludes the proof.

**Remark 12.1.** Under (O2), we have  $|\eta_{\mathcal{H}}^{\textcircled{a}}(x)| \leq M$ , and since  $\eta \in \mathcal{L}^s(\mathcal{E}, M)$ , we obtain  $|\eta(x) - \eta_{\mathcal{H}}^{\textcircled{a}}(x)| \leq 2M$  on the rhs of eq. (23). While on the lhs, it is straightforward that (see [25, Chap. 10])

$$|\eta(x) - \eta_{\mathcal{H}}^{\textcircled{a}}(x)| = |\eta(x) - T_M(\eta_{\mathcal{H}}^{\diamond}(x))| \leq |\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)|.$$

Under (O1), the proof remains unchanged. So that the proof still holds with

$$\bar{M} = \begin{cases} 2M, & \text{under (O2),} \\ M, & \text{otherwise.} \end{cases}$$

□

**Local estimation under (CS1) in the case where  $r \geq 2$ .** The proof follows the same lines as above, except that  $Q_{\mathcal{H}}$  is now a matrix. Let  $x \in \mathcal{A}$  and  $j \in \mathcal{J}_n$ . There exists  $\mathcal{H} \in \mathcal{F}_j$  such that  $x \in \mathcal{H}$ . Let us work on the set  $\{\lambda_{\min}(Q_{\mathcal{H}}) \geq \pi_n^{-1}\}$  on which  $Q_{\mathcal{H}}$  is invertible. On that set, we can write

$$\begin{aligned} |\mathcal{P}_j \eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| &= |\langle \alpha_{\mathcal{H}} - \alpha_{\mathcal{H}}^{\diamond}, \varphi_{\mathcal{H}}(x) \rangle| \\ &= |\langle Q_{\mathcal{H}}^{-1} \cdot \left( \frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}}) \right), \varphi_{\mathcal{H}}(x) \rangle| \\ &\leq \|Q_{\mathcal{H}}^{-1}\|_S \left\| \frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}}) \right\|_{\ell^2(\mathbb{R}^{R^d})} \|\varphi_{\mathcal{H}}(x)\|_{\ell^2(\mathbb{R}^{R^d})} \\ &\leq R^{\frac{d}{2}} 2^{j\frac{d}{2}} \lambda_{\min}(Q_{\mathcal{H}})^{-1} \left\| \frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}}) \right\|_{\ell^2(\mathbb{R}^{R^d})}. \end{aligned}$$

Now, notice that for all  $X_i \in \mathcal{H}$ , we have  $Y_i = \langle \alpha_{\mathcal{H}}, \varphi_{\mathcal{H}}(X_i) \rangle + \mathcal{R}_j \eta(X_i) + \xi_i$ . Write  $\mathcal{R}_{\mathcal{H}} = (\mathcal{R}_j \eta(X_i) \mathbb{1}_{\mathcal{H}}(X_i))_{1 \leq i \leq n}$  and  $\xi_{\mathcal{H}} = (\xi_i \mathbb{1}_{\mathcal{H}}(X_i))_{1 \leq i \leq n}$ . Then, we have,

$$W_{\mathcal{H}} = \left| \frac{B_{\mathcal{H}}^t}{n} \cdot (B_{\mathcal{H}} \cdot \alpha_{\mathcal{H}} - Y_{\mathcal{H}}) \right| = \left| \frac{B_{\mathcal{H}}^t}{n} \cdot (\xi_{\mathcal{H}} + \mathcal{R}_{\mathcal{H}}) \right| \in \mathbb{R}^{R^d}.$$

Thus, a direct application of [Proposition 12.6](#) allows to write, for  $\delta > 2M2^{-js} \max(1, 3\pi_n R^d \mu_{\max})$ ,

$$\begin{aligned} \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| \geq \delta, \lambda_{\min}(Q_{\mathcal{H}}) \geq \pi_n^{-1}) \\ \leq \mathbb{P}(\|W_{\mathcal{H}}\|_{\ell^2(\mathbb{R}^{R^d})} \geq \frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^{\frac{d}{2}}}) \\ \leq R^d \sup_{k \in \mathcal{S}_j(\mathcal{H})} \mathbb{P}\left([W_{\mathcal{H}}]_k \geq \frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right) \\ \leq R^d \Lambda\left(\frac{\delta 2^{-j\frac{d}{2}}}{2\pi_n R^d}\right). \end{aligned}$$

At last, using exactly the same arguments as in the case where  $r = 1$  above, we obtain

$$\begin{aligned} \mathbb{P}(|\eta(x) - \eta_j^{\circledast}(x)| \geq \delta) &\leq \mathbb{P}(|\eta(x) - \eta_{\mathcal{H}}^{\diamond}(x)| \geq \delta, \lambda_{\min}(Q_{\mathcal{H}}) \geq \pi_n^{-1}) \\ &\quad + \mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) < \pi_n^{-1}) \mathbb{1}_{\{\delta \leq \bar{M}\}}, \end{aligned}$$

where  $\bar{M}$  has been defined in [Remark 12.1](#). The term on the lhs has been dealt with above. The term on the rhs is tackled using [Proposition 12.3](#). This concludes the proof.  $\square$

### 12.1.3 Proof of [Theorem 7.2](#)

This result is obtained after a slight modification of [\[35, Proposition 3.4\]](#). In the same way as in the proof of [Theorem 7.1](#), we are brought back to controlling  $\mathbb{E}|\eta_{j^{\circledast}(x)}^{\circledast}(x) - \eta(x)|^p$  for all  $x \in \mathcal{A}$ . To that end, we split this term as follows

$$\begin{aligned} \mathbb{E}|\eta_{j^{\circledast}(x)}^{\circledast}(x) - \eta(x)|^p &= \mathbb{E}|\eta_{j^{\circledast}(x)}^{\circledast}(x) - \eta(x)|^p (\mathbb{1}_{\{j^{\circledast}(x) \leq j_s\}} + \mathbb{1}_{\{j^{\circledast}(x) > j_s\}}) \\ &= I + II. \end{aligned}$$

Let us first deal with  $I$ . Notice that

$$2^{1-p} |\eta_{j^{\textcircled{a}}(x)}^{\textcircled{a}}(x) - \eta(x)|^p \leq |\eta_{j^{\textcircled{a}}(x)}^{\textcircled{a}}(x) - \eta_{j_s}^{\textcircled{a}}(x)|^p + |\eta_{j_s}^{\textcircled{a}}(x) - \eta(x)|^p.$$

The last term is of the good order since

$$\begin{aligned} \mathbb{E} |\eta_{j_s}^{\textcircled{a}}(x) - \eta(x)|^p &\leq C(p) \pi_n^p \max \left( 2^{-j_s s}, \frac{2^{j_s \frac{d}{2}}}{\sqrt{n}} \right)^p \\ &= \frac{C(p)}{(\kappa \log n)^{\frac{p}{2}}} \left( t(n) 2^r n^{-\frac{s}{2s+d}} \right)^p, \end{aligned}$$

according to [Proposition 12.1](#), [eq. \(10a\)](#) and [eq. \(10c\)](#). Regarding the first term, notice that on the event  $\{j^{\textcircled{a}}(x) \leq j_s\}$ , one has got

$$\begin{aligned} |\eta_{j^{\textcircled{a}}(x)}^{\textcircled{a}}(x) - \eta_{j_s}^{\textcircled{a}}(x)| &\leq g(j^{\textcircled{a}}(x), j_s) \leq \sup_{j^{\textcircled{a}} \leq k \leq j_s} g(k, j_s) \\ &\leq g(j_s, j_s) = 2t(n) \frac{2^{j_s \frac{d}{2}}}{\sqrt{n}} \leq 2t(n) 2^r n^{-\frac{s}{2s+d}}, \end{aligned}$$

where we have used [eq. \(10a\)](#) and [eq. \(10c\)](#) and which is of the good order too. Let us now turn to  $II$ . For any two  $j < k$ , we write

$$\mathcal{G}(x, j, k) = \{|\eta_j^{\textcircled{a}}(x) - \eta_k^{\textcircled{a}}(x)| > g(j, k)\}.$$

Write  $\mathcal{J}_n(j) = \{k \in \mathcal{J}_n : k > j\}$ . Notice first that we have the following inclusions

$$\begin{aligned} \{j^{\textcircled{a}}(x) = j\} &\subseteq \bigcup_{k \in \mathcal{J}_n(j-1)} \mathcal{G}(x, j-1, k), \\ \{j^{\textcircled{a}}(x) > j_s\} &= \bigcup_{j \in \mathcal{J}_n(j_s)} \{j^{\textcircled{a}}(x) = j\} \subseteq \bigcup_{j \in \mathcal{J}_n(j_s)} \bigcup_{k \in \mathcal{J}_n(j-1)} \mathcal{G}(x, j-1, k). \end{aligned}$$

Therefore, we can write

$$\begin{aligned} II &\leq \sum_{j \in \mathcal{J}_n(j_s)} \mathbb{E} |\eta_{j^{\textcircled{a}}(x)}^{\textcircled{a}}(x) - \eta(x)|^p \mathbf{1}_{\{j^{\textcircled{a}}(x)=j\}} \\ &\leq \sum_{j \in \mathcal{J}_n(j_s)} \sum_{k \in \mathcal{J}_n(j-1)} \mathbb{E} |\eta_j^{\textcircled{a}}(x) - \eta(x)|^p \mathbf{1}_{\mathcal{G}(x, j-1, k)}. \end{aligned}$$

Now, we notice that

$$|\eta_j^{\textcircled{a}}(x) - \eta_k^{\textcircled{a}}(x)| \leq |\eta_j^{\textcircled{a}}(x) - \eta(x)| + |\eta(x) - \eta_k^{\textcircled{a}}(x)|.$$

So that

$$\begin{aligned}\mathcal{G}(x, j, k) &= \{|\eta_j^\oplus(x) - \eta_k^\oplus(x)| > g(j, k)\} \\ &\subset \left\{|\eta_j^\oplus(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}}t(n)\right\} \cup \left\{|\eta_k^\oplus(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}}t(n)\right\}, \\ \mathbb{P}(\mathcal{G}(x, j, k)) &\leq \mathbb{P}\left(|\eta_j^\oplus(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}}t(n)\right) + \mathbb{P}\left(|\eta_k^\oplus(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}}t(n)\right).\end{aligned}$$

So that a direct application of the Cauchy-Schwarz inequality leads to

$$\mathbb{E}|\eta_j^\oplus(x) - \eta(x)|^p \mathbf{1}_{\mathcal{G}(x, j-1, k)} \leq (\mathbb{E}|\eta_j^\oplus(x) - \eta(x)|^{2p})^{\frac{1}{2}} \mathbb{P}(\mathcal{G}(x, j-1, k))^{\frac{1}{2}}.$$

Now, a direct application of [Proposition 12.1](#) for  $j_s \leq j \leq J$  gets us

$$(\mathbb{E}|\eta_j^\oplus(x) - \eta(x)|^{2p})^{\frac{1}{2}} \leq \sqrt{C(2p)}\pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p \leq \sqrt{C(2p)}(\kappa \log n)^{-\frac{p}{2}}.$$

Besides, notice that for  $j_s \leq j < k \leq J$ , we can apply [Proposition 12.2](#) with  $\kappa \geq \frac{p}{2}C_9^{-1}$  to obtain

$$\mathbb{P}\left(|\eta_j^\oplus(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}}t(n)\right) \vee \mathbb{P}\left(|\eta_k^\oplus(x) - \eta(x)| > \frac{2^{k\frac{d}{2}}}{\sqrt{n}}t(n)\right) \leq 5R^{2d}n^{-\frac{p}{2}}.$$

To conclude the proof, it remains to notice that  $\#\mathcal{J}_n \leq \log n$  and remark that the multiplicative constant in the upper-bound of [Theorem 7.2](#) is indeed smaller than, say, 5 for  $n$  large enough.  $\square$

#### 12.1.4 A few useful Propositions and Lemmas

**Proposition 12.1.** *Fix  $r \in \mathbb{N}$  and assume we are under (CS1) and  $(\mathbf{H}_s^r)$ . Then, For any  $x \in \mathcal{A}$  and  $j \in \mathcal{J}_n$ , one has got*

$$\mathbb{E}[|\eta(x) - \eta_j^\oplus(x)|^p] \leq C(p)\pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p,$$

where

$$C(p) = 3^p M^p \max(1, R^{2d}\mu_{\max})^p + C_5(r, d, p, \mu_{\max}; K, \sigma) + 2M^p R^{2d},$$

and  $C_5$  is made explicit in the proof at [eq. \(24\)](#).

*Proof.* For any  $x \in \mathcal{A}$ , take  $\delta = 3M2^{-js} \max(1, 3\pi_n R^d \mu_{\max})$ . Notice first that  $\max(1, 3\pi_n R^d \mu_{\max}) \leq \pi_n \max(1, 3R^d \mu_{\max})$  since, by construction,  $\pi_n^{-1} \leq 1$  in any case. Now, write

$$\begin{aligned}\mathbb{E}[|\eta(x) - \eta_j^\oplus(x)|^p] &= \int_{\mathbb{R}^+} pt^{p-1} \mathbb{P}(|\eta(x) - \eta_j^\oplus(x)| \geq t) dt \\ &\leq \delta^p + \int_{\delta}^{+\infty} pt^{p-1} \mathbb{P}(|\eta(x) - \eta_j^\oplus(x)| \geq t) dt.\end{aligned}$$

As  $\delta$  has been fixed, we only need to tackle the rhs above, which we will denote by  $II$ . Using [Theorem 7.1](#), we can write

$$II \leq 2R^{2d} \exp \left( -n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1}} \right) \int_0^M pt^{p-1} dt \\ + R^d \int_0^\infty pt^{p-1} \Lambda \left( \frac{t2^{-j\frac{d}{2}}}{2\pi_n R^d} \right) dt.$$

Denote by  $II_1$  and  $II_2$  the lhs and rhs terms above, respectively. Now, recall that  $j \leq J$ , where  $2^{Jd} \leq nt(n)^{-2}$  and  $t(n)^2 = \kappa\pi_n^2 \log n$ . Therefore, as soon as

$$\kappa \geq \frac{p}{2} \left( 2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}\pi_n^{-1} \right),$$

we have  $II_1 \leq 2M^p R^{2d} n^{-\frac{p}{2}}$ . Let us now turn to  $II_2$ . Assume first that we are working under the bounded noise assumption, **(N1)**. In that case, we have

$$II_2 \leq 2R^d \int_0^\infty pt^{p-1} \exp \left( -\frac{n2^{-jd}t^2\pi_n^{-2}}{64K^2R^{2d}\mu_{\max} + 8KR^d\pi_n^{-1}t} \right) dt \\ \leq C_2(r, d, p, \mu_{\max}, K) \left( \pi_n \frac{2^{j\frac{d}{2}}}{\sqrt{n}} \right)^p.$$

where the last inequality results from the change of variable  $u = \sqrt{n}2^{-j\frac{d}{2}}\pi_n^{-1}t$  together with the fact that  $2^{jd} \leq n$  and we have written

$$C_2 := 2R^d \int_0^\infty pt^{p-1} \exp \left( -\frac{t^2}{64K^2R^{2d}\mu_{\max} + 8KR^d t} \right) dt.$$

Assume now that we are working under the Gaussian noise assumption **(N2)**. In that case, we have

$$II_2 \leq R^d \int_0^\infty pt^{p-1} \left( 1 \wedge \left\{ \frac{2\sigma R^{\frac{d}{2}}(4R^d\mu_{\max} + 2t\pi_n^{-1})^{\frac{1}{2}}}{t\pi_n^{-1}2^{-j\frac{d}{2}}\sqrt{2\pi n}} \right. \right. \\ \left. \left. \exp \left( -\frac{n2^{-jd}\pi_n^{-2}t^2\sigma^{-2}}{4R^{2d}\mu_{\max} + 2R^d\pi_n^{-1}t} \right) \right\} \right) dt \\ + 2R^d \int_0^\infty pt^{p-1} \exp \left( -\frac{n2^{-jd}\pi_n^{-2}t^2}{8R^{2d}\mu_{\max} + \frac{8}{3}R^d\pi_n^{-1}t} \right) dt.$$

Denote by  $II_3$  and  $II_4$  the first and second term, respectively. They can both be handled in the exact same way as  $II_2$ , which leads to

$$II_4 \leq C_4(r, d, p, \mu_{\max}) \left( \pi_n \frac{2^{j\frac{d}{2}}}{\sqrt{n}} \right)^p,$$

where we have written

$$C_4 := 2R^d \int_0^\infty pt^{p-1} \exp\left(-\frac{t^2}{8R^{2d}\mu_{\max} + \frac{8}{3}R^d t}\right) dt,$$

and

$$II_3 \leq C_3(r, d, p, \mu_{\max}, \sigma) \left( \pi_n \frac{2^{j\frac{d}{2}}}{\sqrt{n}} \right)^p,$$

where we have written

$$C_3 := R^d \int_0^\infty pt^{p-1} \left( 1 \wedge \left\{ \frac{2\sigma R^{\frac{d}{2}}(4R^d\mu_{\max} + 2t)^{\frac{1}{2}}}{t\sqrt{2\pi}} \exp\left(-\frac{t^2\sigma^{-2}}{4R^{2d}\mu_{\max} + 2R^d t}\right) \right\} \right) dt.$$

To conclude, let us write

$$C_5(r, d, p, \mu_{\max}; K, \sigma) = \begin{cases} C_2(r, d, p, \mu_{\max}, K) & \text{under (N1)} \\ C_3(r, d, p, \mu_{\max}, \sigma) + C_4(r, d, p, \mu_{\max}) & \text{under (N2)} \end{cases} \quad (24)$$

Therefore, we ultimately obtain

$$\begin{aligned} \mathbb{E}[|\eta(x) - \eta_j^\circ(x)|^p] &\leq (3^p M^p \max(1, 3R^d \mu_{\max})^p + C_5 + 2M^p R^{2d}) \\ &\quad \pi_n^p \max\left(2^{-js}, \frac{2^{j\frac{d}{2}}}{\sqrt{n}}\right)^p, \end{aligned}$$

which concludes the proof.  $\square$

**Proposition 12.2.** *Fix  $r$  in  $\mathbb{N}$  and assume we are under (CS1) and  $(\mathbf{H}_s^r)$ . This means in particular that  $s \in (0, r)$ . Let  $j$  be such that  $j_s \leq j \leq J$ . Let  $t(n)^2 = \kappa \pi_n^2 \log n$ , and define*

$$C_9(r, d, \mu_{\max}, \pi_n; K, \sigma) := \begin{cases} C_6(r, d, \mu_{\max}, K, \pi_n), & \text{under (N1)} \\ C_6(r, d, \mu_{\max}, \sigma, \pi_n), & \text{under (N2)} \end{cases},$$

where  $C_6$  is defined in [eq. \(25\)](#) below. Then we have, for  $n$  large enough,

$$\mathbb{P}\left(|\eta_j^\circ(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n)\right) \leq 5R^{2d} n^{-\kappa C_9}.$$

*Proof.* The proof relies on a direct application of [Theorem 7.1](#). Write  $C_0 = 2M \max(1, 3\pi_n R^d \mu_{\max})$  and notice indeed that the theorem applies since for  $j \geq j_s$ , we get  $2^{j\frac{d}{2}} n^{-\frac{1}{2}} \geq 2^{-(r+\frac{d}{2})} 2^{-j_s s}$  (see [eq. \(10b\)](#)) and, as soon as  $n$  is large enough, we have  $t(n) \geq 2^{r+\frac{d}{2}} C_0$ . This leads us to

$$\mathbb{P} \left( |\eta_j^{\oplus}(x) - \eta(x)| > \frac{2^{j\frac{d}{2}}}{\sqrt{n}} t(n) \right) \leq 2R^{2d} \exp \left( -n2^{-jd} \frac{\pi_n^{-2}}{2\mu_{\max} R^{4d} + \frac{4}{3} R^{2d} \pi_n^{-1}} \right) + R^d \Lambda \left( \frac{t(n)}{2\pi_n R^d \sqrt{n}} \right).$$

Let us denote the first term by  $I$  and the second one by  $II$ .  $I$  is easily tackled noticing that for  $j \leq J$ ,  $n2^{-jd} \geq n2^{-Jd} \geq t(n)^2 = \kappa \pi_n^2 \log n$ . So that, we obtain  $I \leq 2R^{2d} n^{-\kappa C_6}$ , where we have written

$$C_6(r, d, \mu_{\max}, K, \pi_n) := \frac{\min(1, K^{-2})}{64\mu_{\max} R^{2d} + 8R^d \pi_n^{-1}}. \quad (25)$$

Let us now turn to  $II$ . Assume first we work under **(N1)**. Then we can write

$$II \leq 2R^d \exp \left( - \frac{t(n)^2 \pi_n^{-2}}{64R^{2d} K^2 \mu_{\max} + 8R^d K \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}} \right).$$

Notice first that  $2^{j\frac{d}{2}} t(n) \leq \sqrt{n}$ . Therefore, we obtain  $II \leq 2R^d n^{-\kappa C_6}$ . Assume now that we work under **(N2)**. In that case, we obtain

$$\begin{aligned} II &\leq R^d \left( 1 \wedge \left\{ \frac{2R^{\frac{d}{2}} \sigma (4R^d \mu_{\max} + 2\pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}})^{\frac{1}{2}}}{t(n) \pi_n^{-1} \sqrt{2\pi}} \right. \right. \\ &\quad \left. \left. \exp \left( - \frac{t(n)^2 \pi_n^{-2} \sigma^{-2}}{4R^{2d} \mu_{\max} + 2R^d \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}} \right) \right\} \right) \\ &\quad + 2R^d \exp \left( - \frac{t(n)^2 \pi_n^{-2}}{8R^{2d} \mu_{\max} + \frac{8}{3} R^d \pi_n^{-1} \frac{2^{j\frac{d}{2}} t(n)}{\sqrt{n}}} \right). \end{aligned}$$

We proceed exactly as under **(N1)**. So that we obtain  $II \leq C_7 n^{-\kappa C_8}$ , where

$$\begin{aligned} C_8(r, d, \mu_{\max}, \sigma, \pi_n) &:= \frac{\min(1, \sigma^{-2})}{4R^{2d} \mu_{\max} + 2R^d \pi_n^{-1}}, \\ C_7(r, d, \mu_{\max}, \sigma, \pi_n, t(n)) &= R^d \frac{2R^{\frac{d}{2}} \sigma (4R^d \mu_{\max} + 2\pi_n^{-1})^{\frac{1}{2}}}{t(n) \pi_n^{-1} \sqrt{2\pi}} + 2R^d. \end{aligned}$$

So that  $C_7 \leq 3R^d$  for  $n$  large enough. Notice finally that  $C_8(r, d, \mu_{\max}, t, \pi_n) \geq C_6(r, d, \mu_{\max}, t, \pi_n)$ . This concludes the proof.  $\square$

**Proposition 12.3.** Fix an integer  $r \geq 2$  and assume we are under (CS1). Let  $x \in \mathcal{A}$  and  $j \in \mathcal{J}_n$ . By construction, there exists  $\mathcal{H} \in \mathcal{F}_j$  such that  $x \in \mathcal{H}$ . Recall besides that  $\#\mathcal{S}_j(\mathcal{H}) = R^d$ , where  $R = 2r - 1$  is obviously independent of both  $x$  and  $j$ . Write  $\|\cdot\| = \|\cdot\|_{\ell_2(\mathbb{R}^{R^d})}$  and assume there exists a strictly positive constant  $g_{\min}$  independent of  $x$  and  $j$  such that

$$\lambda_{\min}(\mathbb{E}Q_{\mathcal{H}}) = \min_{u \in \mathbb{R}^{R^d}: \|u\|=1} \langle u, \mathbb{E}Q_{\mathcal{H}}u \rangle \geq g_{\min}. \quad (26)$$

Then, for any real number  $t$  such that  $0 < t \leq \frac{g_{\min}}{2}$ , we have

$$\mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) \leq t) \leq 2R^{2d} \exp \left( -n2^{-jd} \frac{t^2}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}t} \right).$$

*Proof.* Under the assumption described in eq. (26), we get

$$\begin{aligned} \lambda_{\min}(Q_{\mathcal{H}}) &\geq \min_{u \in \mathbb{R}^{R^d}: \|u\|=1} \langle u, \mathbb{E}Q_{\mathcal{H}}u \rangle + \min_{u \in \mathbb{R}^{R^d}: \|u\|=1} \langle u, (Q_{\mathcal{H}} - \mathbb{E}Q_{\mathcal{H}})u \rangle \\ &\geq 2t - \sum_{\nu, \nu' \in \mathcal{S}_j(\mathcal{H})} |[Q_{\mathcal{H}}]_{\nu, \nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu, \nu'}|. \end{aligned}$$

Write  $T_i = \varphi_{j, \nu}(X_i)\varphi_{j, \nu'}(X_i)\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j, \nu}(X)\varphi_{j, \nu'}(X)\mathbb{1}_{\mathcal{H}}(X)$ , so that  $\mathbb{E}T_i = 0$ ,  $\text{Var}T_i \leq \mu_{\max}2^{jd}$  and  $|T_i| \leq 2^{jd+1}$ . A direct application of Bernstein inequality for any  $\delta > 0$  leads to

$$\begin{aligned} &\mathbb{P}(|[Q_{\mathcal{H}}]_{\nu, \nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu, \nu'}| \geq \delta) \\ &= \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \varphi_{j, \nu}(X_i)\varphi_{j, \nu'}(X_i)\mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E}\varphi_{j, \nu}(X)\varphi_{j, \nu'}(X)\mathbb{1}_{\mathcal{H}}(X)\right| \geq \delta\right) \\ &\leq 2 \exp \left( -\frac{n2^{-jd}\delta^2}{2\mu_{\max} + \frac{4}{3}\delta} \right). \end{aligned}$$

To conclude, we write

$$\begin{aligned} \mathbb{P}(\lambda_{\min}(Q_{\mathcal{H}}) \leq t) &\leq \mathbb{P}\left(\sum_{\nu, \nu' \in \mathcal{S}_j(\mathcal{H})} |[Q_{\mathcal{H}}]_{\nu, \nu'} - [\mathbb{E}Q_{\mathcal{H}}]_{\nu, \nu'}| \geq t\right) \\ &\leq 2R^{2d} \exp \left( -n2^{-jd} \frac{t^2}{2\mu_{\max}R^{4d} + \frac{4}{3}R^{2d}t} \right). \end{aligned}$$

□

**Proposition 12.4.** Fix an integer  $r \geq 2$  and assume we are under (CS1). For any  $x \in \mathcal{A}$  and  $j \in \mathcal{J}_n$ , we denote by  $\mathcal{H}$  the unique hypercube of  $\mathcal{F}_j$  such that  $x \in \mathcal{H}$ . Then, there exists a strictly positive absolute constant  $g_{\min}$  independent of both  $x$  and  $j$  such that, for all  $j \in \mathcal{J}_n$  and all  $x \in \mathcal{A}$ , we have  $\lambda_{\min}(\mathbb{E}Q_{\mathcal{H}}) \geq g_{\min} > 0$ .

*Proof.* For any  $u \in \mathbb{R}^{R^d}$  such that  $\|u\|_{\ell_2(\mathbb{R}^{R^d})} = 1$ , we can write

$$\langle u, \mathbb{E}Q_{\mathcal{H}} \cdot u \rangle = \int_{\mathcal{A}} \left( \sum_{\nu \in \mathcal{S}_j(\mathcal{H})} u_{\nu} \varphi_{j, \nu}(w) \mathbb{1}_{\mathcal{H}}(w) \right)^2 \mu(w) dw$$



$$\geq \mu_{\min} \int_{\mathcal{H}} \left( \sum_{\nu \in \mathcal{S}_j(\mathcal{H})} u_{\nu} \varphi_{j,\nu}(w) \right)^2 dw, \quad (27)$$

$$= \mu_{\min} \int_{[0,1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_{\nu} \varphi_{\nu}(w) \right)^2 dw, \quad (28)$$

where  $\mathfrak{S}$  has been defined in eq. (15) and the last equality results from the fact that the value of the integral on the rhs of eq. (27) is invariant with  $\mathcal{H}$ , meaning that for any  $\mathcal{H} \in \mathcal{F}_j$  and  $u \in \mathbb{R}^{R^d}$ ,

$$\begin{aligned} & \int_{\mathcal{H}} \left( \sum_{\nu \in \mathcal{S}_j(\mathcal{H})} u_{\nu} \varphi_{j,\nu}(w) \right)^2 dw \\ &= \int_{2^{-j}[0,1]^d} \left( \sum_{\nu \in \mathcal{S}_j(2^{-j}[0,1]^d)} u_{\nu} \varphi_{j,\nu}(w) \right)^2 dw \\ &= \int_{[0,1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_{\nu} \varphi_{\nu}(w) \right)^2 dw. \end{aligned}$$

Let us denote by  $\mathbb{S}^{R^d-1}$  the unit-sphere of  $\mathbb{R}^{R^d}$ . As detailed in the proof of Lemma 12.1, the map

$$u \in \mathbb{S}^{R^d-1} \mapsto \int_{[0,1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_{\nu} \varphi_{\nu}(w) \right)^2 dw,$$

is absolutely continuous with respect to  $u$  on the compact subset  $\mathbb{S}^{R^d-1}$  of  $\mathbb{R}^{R^d}$ . It therefore reaches its minimum at some point  $u^* \in \mathbb{S}^{R^d-1}$ . It is a direct consequence of the **local linear independence property** of the scaling functions  $(\varphi_k)$  (see Proposition 12.8) that

$$\int_{[0,1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_{\nu}^* \varphi_{\nu}(w) \right)^2 dw = c_{\min} > 0,$$

where  $c_{\min}$  is a constant that is both independent from  $x$  and  $j$ . This concludes the proof with  $g_{\min} = \mu_{\min} c_{\min}$ .  $\square$

**Proposition 12.5.** *Fix  $r = 1$  and assume we are under (CS1). Let  $x \in \mathcal{A}$ ,  $j \in \mathcal{J}_n$  and  $\mathcal{H} \in \mathcal{F}_j$  such that  $x \in \mathcal{H}$ . Then, for any real number  $t$  such that  $0 < t \leq \frac{g_{\min}}{2}$ , we have  $2^{jd} \frac{n_{\mathcal{H}}}{n} \geq t$  on the event  $\Omega_{j,\nu(\mathcal{H})}^c(t)$ , where  $\Omega_{j,\nu(\mathcal{H})}(t) := \{\frac{1}{n} \sum_{i=1}^n \varphi_{j,\nu(\mathcal{H})}(X_i)^2 < t\}$ . In addition, we have got*

$$\mathbb{P}(\Omega_{j,\nu(\mathcal{H})}(t)) \leq 2 \exp \left( -n 2^{-jd} \frac{t^2}{2\mu_{\max} + \frac{4}{3}t} \right).$$

Recall that the constant  $g_{\min}$  has been defined in Proposition 12.4.

*Proof.* Notice that

$$\int \varphi_{j,\nu(\mathcal{H})}(w)^2 \mu(w) dw \geq \mu_{\min} \geq 2t. \quad (29)$$

On  $\Omega_{j,\nu(\mathcal{H})}^c(t)$ , we have

$$2^{jd} \frac{n\mathcal{H}}{n} \geq \frac{1}{n} \sum_{i=1}^n \varphi_{j,\nu(\mathcal{H})}(X_i)^2 \geq t.$$

For any  $\delta > 0$ , we write

$$\mathcal{C}_{j,k}(\delta) = \left\{ \left| \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i)^2 - \mathbb{E} \varphi_{j,k}(X)^2 \right| \leq \delta \right\}. \quad (30)$$

Notice now that

$$\Omega_{j,\nu(\mathcal{H})}(t) \subset \left\{ \frac{1}{n} \sum_{i=1}^n \varphi_{j,\nu(\mathcal{H})}(X_i)^2 - \mathbb{E} \varphi_{j,\nu(\mathcal{H})}(X)^2 < -t \right\} \subset \mathcal{C}_{j,\nu(\mathcal{H})}^c(t),$$

where  $\mathcal{C}_{j,\nu(\mathcal{H})}(t)$  is defined in [eq. \(30\)](#). Finally, it is straightforward that  $\mathcal{C}_{j,\nu(\mathcal{H})}^c(t)$  has very small probability thanks to Bernstein inequality (see [\[27, p. 165\]](#) for example). Write indeed  $T_i = \varphi_{j,\nu(\mathcal{H})}(X_i)^2 - \mathbb{E} \varphi_{j,\nu(\mathcal{H})}(X)^2$ . Then  $\mathbb{E} T_i = 0$ ,  $\text{Var} T_i \leq \mu_{\max} 2^{jd}$  and  $|T_i| \leq 2^{jd+1}$ . Thus, we can write

$$\mathbb{P}(\mathcal{C}_{j,\nu(\mathcal{H})}^c(t)) \leq 2 \exp \left( -n 2^{-jd} \frac{t^2}{2\mu_{\max} + \frac{4}{3}t} \right),$$

which concludes the proof.  $\square$

**Proposition 12.6.** *Let  $(X_i)_{i=1,\dots,n}$  and  $(\xi_i)_{i=1,\dots,n}$  be sequences of independent random variables such that  $\mathbb{E}(\xi|X) = 0$ . Take any  $j \geq j_r$ . Moreover, assume we are given a function  $\mathcal{R}_j(\cdot)$  such that  $\|\mathcal{R}_j(\cdot)\|_{\mathbb{L}_\infty(\mathcal{E},\lambda)} \leq M 2^{-js}$ , a subset  $\mathcal{H}$  of  $\mathcal{E}$  and a scaling function  $\varphi_{j,k}$ . Write*

$$W_{j,k} = \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i) \mathbf{1}_{\mathcal{H}}(X_i) (\mathcal{R}_j(X_i) + \xi_i),$$

and define

$$\Lambda(\delta) = \begin{cases} 2 \exp \left( -\frac{n\delta^2}{18K^2\mu_{\max} + 4K2^{j\frac{d}{2}}\delta} \right), & \text{under (N1)} \\ 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp \left( -\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta} \right) \right\} \\ + 2 \exp \left( -\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta} \right), & \text{under (N2)} \end{cases}$$

Then, for all  $\delta > 3\mu_{\max}M2^{-j(s+\frac{d}{2})}$ , we have

$$\mathbb{P}(|W_{j,k}| \geq \delta) \leq \Lambda(\delta).$$

*Proof.* Notice indeed that

$$\begin{aligned} W_{j,k} &\leq \left| \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i) \xi_i \mathbb{1}_{\mathcal{H}}(X_i) \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i) \mathcal{R}_j(X_i) \mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E} \varphi_{j,k}(X) \mathcal{R}_j(X) \mathbb{1}_{\mathcal{H}}(X) \right| \\ &\quad + |\mathbb{E} \varphi_{j,k}(X) \mathcal{R}_j(X) \mathbb{1}_{\mathcal{H}}(X)| \\ &= I + II + III. \end{aligned}$$

So that we can write

$$\mathbb{P}(|W_{j,k}| \geq \delta) \leq \mathbb{P}(I \geq \delta/3) + \mathbb{P}(II \geq \delta/3) + \mathbb{P}(III \geq \delta/3).$$

Now it is enough to notice that

$$\begin{aligned} III &\leq \int |\varphi_{j,k}(w) \mathcal{R}_j(w)| \mathbb{1}_{\mathcal{H}}(w) \mu(w) dw \\ &\leq \mu_{\max} \int_{\mathcal{E}} |\varphi_{j,k}(w) \mathcal{R}_j(w)| dw \\ &\leq \mu_{\max} \|\varphi_{j,k}\|_{\mathbb{L}_1(\mathcal{E}, \lambda)} \|\mathcal{R}_j\|_{\mathbb{L}_{\infty}(\mathcal{E}, \lambda)} \\ &\leq \mu_{\max} M 2^{-j(s+\frac{d}{2})}. \end{aligned}$$

So that  $\mathbb{P}(III \geq \delta/3) = 0$  as soon as  $\delta > 3\mu_{\max}M2^{-j(s+\frac{d}{2})}$ .

Now, turn to  $II$  and write  $II = |\sum T_i/n|$  with  $T_i = \varphi_{j,k}(X_i) \mathcal{R}_j(X_i) \mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E} \varphi_{j,k}(X) \mathcal{R}_j(X) \mathbb{1}_{\mathcal{H}}(X)$ . Obviously  $\mathbb{E} T_i = 0$ ,  $\text{Var} T_i \leq \mathbb{E} (\varphi_{j,k}(X) \mathcal{R}_j(X) \mathbb{1}_{\mathcal{H}}(X))^2 \leq \mu_{\max} M^2 2^{-2js}$  and  $|T_i| \leq M 2^{-js} 2^{j\frac{d}{2}+1}$ . So that we can apply Bernstein inequality to get

$$\mathbb{P}(II \geq \delta/3) \leq 2 \exp \left( - \frac{n 2^{2js} \delta^2}{18 \mu_{\max} M^2 + 4 M 2^{j\frac{d}{2}} 2^{js} \delta} \right).$$

And finally, turn to  $III$ . Assume first that the noise  $\xi$  is bounded by  $K$ . We have obviously  $\mathbb{E} \varphi_{j,k}(X_i) \xi_i \mathbb{1}_{\mathcal{H}}(X_i) = 0$ ,  $\text{Var}(\varphi_{j,k}(X_i) \xi_i \mathbb{1}_{\mathcal{H}}(X_i)) \leq K^2 \mu_{\max}$  and  $|\varphi_{j,k}(X_i) \xi_i \mathbb{1}_{\mathcal{H}}(X_i)| \leq K 2^{j\frac{d}{2}+1}$ , so that

$$\mathbb{P}(I \geq \delta/3) \leq 2 \exp \left( - \frac{n \delta^2}{18 K^2 \mu_{\max} + 4 K 2^{j\frac{d}{2}} \delta} \right).$$

Now, it is enough to notice that for all  $s > 0$  and  $j$  such that  $j \geq \frac{1}{s} \log_2 \frac{M}{K}$  (which becomes a constraint for  $K < M$  only),

$$\frac{n 2^{2js} \delta^2}{18 \mu_{\max} M^2 + 4 M 2^{j\frac{d}{2}} 2^{js} \delta} \geq \frac{n \delta^2}{18 K^2 \mu_{\max} + 4 K 2^{j\frac{d}{2}} \delta},$$

which concludes the proof under **(N1)**. When  $j \geq \frac{1}{s} \log_2 3M$ , the conclusion under **(N2)** is a direct consequence of [Proposition 12.7](#).  $\square$

**Proposition 12.7.** *Let  $\varphi_{j,k}$  be a scaling function and  $\mathcal{H}$  a subset of  $\mathcal{E}$ . Define*

$$I = \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i) \xi_i \mathbb{1}_{\mathcal{H}}(X_i).$$

*Assume now that the noise  $\xi$  is conditionally Gaussian, that is we are under **(N2)**. Then, we notice that, conditionally on  $X_1, \dots, X_n$ ,  $I \sim \Phi(0, \sigma \rho_{j,k} / \sqrt{n})$ , where  $\rho_{j,k}^2 = n^{-1} \sum_{i=1}^n \varphi_{j,k}(X_i)^2 \mathbb{1}_{\mathcal{H}}(X_i)$ . Then, for all  $\delta > 0$ , one can write*

$$\begin{aligned} \mathbb{P}(|I| \geq \delta) &\leq 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right) \right\} \\ &\quad + 2 \exp\left(-\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right). \end{aligned}$$

*Proof.* For any  $\delta > 0$ , we write

$$\mathcal{C}_{j,k}(\delta) = \left\{ \left| \frac{1}{n} \sum_{i=1}^n \varphi_{j,k}(X_i)^2 \mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E} \varphi_{j,k}(X)^2 \mathbb{1}_{\mathcal{H}}(X) \right| \leq \delta \right\}.$$

Notice first that

$$\mathcal{C}_{j,k}(2^{j\frac{d}{2}}\delta) \subset \{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}.$$

So that

$$\begin{aligned} \mathbb{1}_{\{|I| \geq \delta\}} &\leq \mathbb{1}_{\{|I| \geq \delta\}} \mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}} + \mathbb{1}_{\{|I| \geq \delta\}} \mathbb{1}_{\mathcal{C}_{j,k}^c(2^{j\frac{d}{2}}\delta)} \\ &\leq \mathbb{1}_{\{|I| \geq \delta\}} \mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}} + \mathbb{1}_{\mathcal{C}_{j,k}^c(2^{j\frac{d}{2}}\delta)}. \end{aligned}$$

The first term is handled thanks to a regular Gaussian tail inequality. Notice indeed that

$$\begin{aligned} \mathbb{P}(|I| \geq \delta | X_1, \dots, X_n) \mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}} \\ \leq 1 \wedge \left\{ \frac{2\rho_{j,k}\sigma}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2}{\rho_{j,k}^2\sigma^2}\right) \right\} \mathbb{1}_{\{\rho_{j,k}^2 \leq \mu_{\max} + 2^{j\frac{d}{2}}\delta\}} \\ \leq 1 \wedge \left\{ \frac{2\sigma(\mu_{\max} + 2^{j\frac{d}{2}}\delta)^{\frac{1}{2}}}{\delta\sqrt{2\pi n}} \exp\left(-\frac{n\delta^2\sigma^{-2}}{\mu_{\max} + 2^{j\frac{d}{2}}\delta}\right) \right\}. \end{aligned}$$

In addition, notice that  $\mathbb{E} \varphi_{j,k}(X)^4 \mathbb{1}_{\mathcal{H}}(X) \leq \mu_{\max} 2^{jd}$  and  $|\varphi_{j,k}(X)^2 \mathbb{1}_{\mathcal{H}}(X_i) - \mathbb{E} \varphi_{j,k}(X)^2 \mathbb{1}_{\mathcal{H}}(X)| \leq 2^{jd+1}$ , so that a direct application of Bernstein inequality leads to

$$\mathbb{P}(\mathcal{C}_{j,k}(2^{j\frac{d}{2}}\delta)^c) \leq 2 \exp\left(-\frac{n2^{jd}\delta^2}{2^{jd}(2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta)}\right) = 2 \exp\left(-\frac{n\delta^2}{2\mu_{\max} + \frac{4}{3}2^{j\frac{d}{2}}\delta}\right),$$

which concludes the proof.  $\square$

## 12.2 Proof of the upper-bound results under (CS2)

Recall that under (CS2), we work with a sample of size  $2n$  split into two pieces denoted by  $\mathcal{D}_n$  and  $\mathcal{D}'_n$ . As detailed previously, similar results as the ones described in [Section 7](#), [Section 8](#) and [Section 12.1.4](#) are still valid with  $\eta^{\star}$  under (CS2). They in fact all stem from [Theorem 9.1](#). The proofs remain for the most part unchanged, with  $\mathcal{J}_n$  redefined as  $\mathcal{J}_n = \{j_s, j_s + 1, \dots, J - 1, J\}$  where  $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$ ,  $\eta^{\star}$  in place of  $\eta^{\oplus}$ ,  $\tilde{X}_i$  in place of  $X_i$  (where we have written  $\tilde{u} = u - X'_{i_x} + 2^{-j-1}$ ), and  $\mathcal{H}_0$  in place of  $\mathcal{H}$ . The sole differences appear in the proofs of [Theorem 9.1](#), [Proposition 12.4](#) and [Proposition 12.5](#). Let us start with the proof of [Theorem 9.1](#).

*Proof of Theorem 9.1.* Assume we are under (CS2) and want to control the probability of deviation of  $\eta_j^{\star}(x)$  from  $\eta(x)$  at a point  $x \in \mathcal{A}$ , for some  $j \in \mathcal{J}_n$ . Recall that  $\mathcal{H}_0(x)$  stands for the cell  $\mathcal{H}_0 = 2^{-j}[0, 1]^d$  centered in  $x$  at level  $j$ , that is  $\mathcal{H}_0(x) = x - 2^{-j-1} + 2^{-j}[0, 1]^d$  and denote by  $\mathcal{O}_x$  the event

$$\mathcal{O}_x = \{\#\{i : X'_i \in \mathcal{H}_0(x)\} \geq 1\}.$$

We can write

$$\begin{aligned} \mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta) &= \mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta, \mathcal{O}_x) \\ &\quad + \mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta, \mathcal{O}_x^c). \end{aligned}$$

Focus first on what happens on the event  $\mathcal{O}_x^c$ . The last term can be controlled easily since the probability that no single design point  $X'_i$  of  $\mathcal{D}'_n$  belongs to  $\mathcal{H}_0(x)$  decreases exponentially fast with  $n$ . Notice indeed that, under (CS2),

$$\begin{aligned} \mathbb{P}(\mathcal{O}_x^c) &= (\mathbb{P}(X'_1 \notin \mathcal{H}_0(x)))^n \\ &= (1 - \mathbb{P}(X'_1 \in \mathcal{H}_0(x)))^n \\ &= \left(1 - \int_{\mathcal{A} \cap \mathcal{H}_0(x)} \mu(w) dw\right)^n \\ &\leq (1 - \mu_{\min} 2^{-jd} \lambda(2^j(\mathcal{A} - x) \cap [-2^{-1}, 2^{-1}]^d))^n \\ &\leq (1 - \mu_{\min} 2^{-jd} \min(2\mathbf{m}_0, 2^{-1})^d)^n \\ &\leq \exp(-\mu_{\min} \min(2\mathbf{m}_0, 2^{-1})^d n 2^{-jd}), \end{aligned}$$

where the before last inequality is a direct consequence of (S2) and the last one comes from the fact that for any  $x \in [0, 1]$ ,  $\ln(1 - x) \leq -x$ . Now, recall that  $\eta_j^{\star}(x) = 0$  on  $\mathcal{O}_x^c$  and  $|\eta(x)| \leq M$  since  $\eta \in \mathcal{L}^s(\mathbb{R}^d, M)$ . So that we obtain

$$\mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta, \mathcal{O}_x^c) \leq \exp(-\mu_{\min} \min(2\mathbf{m}_0, 2^{-1})^d n 2^{-jd}) \mathbf{1}_{\{\delta \leq M\}},$$

which is smaller than the first term in the upper-bound of [Theorem 9.1](#). Now focus on what happens on the event  $\mathcal{O}_x$ . We can write

$$\begin{aligned} \mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta, \mathcal{O}_x) &= \mathbb{P}(\mathcal{O}_x) \mathbb{E}[\mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta | X'_{i_x}) | \mathcal{O}_x] \\ &\leq \mathbb{E}[\mathbb{P}(|\eta(x) - \eta_j^{\star}(x)| \geq \delta | X'_{i_x}) | \mathcal{O}_x]. \end{aligned}$$

Therefore, it is enough to control the probability of deviation of  $\eta_j^\star(x)$  from  $\eta(x)$  on  $\mathcal{O}_x$ , conditionally on  $X'_{i_x}$ . It is controlled in exactly the same way as the probability of deviation of  $\eta_j^\circ(x)$  from  $\eta(x)$  under **(CS1)**, except that we now work with conditional probabilities and expectations with respect to  $X'_{i_x}$ . Interestingly, the random variable  $X'_{i_x}$  is independent of the points of  $\mathcal{D}_n$  since it is built upon the design points  $(X'_i)$  of  $\mathcal{D}'_n$  which are themselves independent of the points of  $\mathcal{D}_n$ . This is a key feature that makes theoretical computations tractable under **(CS2)** and allows to handle  $\eta_j^\star$  in a similar way as  $\eta_j^\circ$  under **(CS1)**. As announced above, [Proposition 12.4](#) and [Proposition 12.5](#) are the sole results that are not obviously true under **(CS2)**. However we show below that they can indeed be extended to setting **(CS2)** without much trouble. Ultimately, this proves that, on the event  $\mathcal{O}_x$  and conditionally on  $X'_{i_x}$ , the probability of deviation of  $\eta_j^\star(x)$  from  $\eta(x)$  verifies [Theorem 7.1](#). So that finally, it remains to put everything together to obtain the results announced in [Theorem 9.1](#), which concludes the proof.  $\square$

Let us now show how the proofs of [Proposition 12.4](#) and [Proposition 12.5](#) can be extended to setting **(CS2)**, thanks to the **local linear independence property** of the scaling functions (see [Proposition 12.8](#)) and a compactness argument.

*Proof of [Proposition 12.4](#) under **(CS2)**.* The proof of [Proposition 12.4](#) under **(CS1)** breaks down under **(CS2)** at boundary points of  $\mathcal{A}$ . [Eq. \(28\)](#) is indeed not valid anymore under **(CS2)** since  $\mu$  eventually cancels on the cell  $\mathcal{H}_0(X'_{i_x})$  centered in  $X'_{i_x}$ . Recall however that  $\mathcal{S}_j(\mathcal{H}_0) = \mathfrak{S} = \{\nu \in \mathbb{Z}^d : 2^{-1} \in \text{Supp}\varphi_\nu\}$  is a set of indexes which is independent of both  $x$  and  $j$ . Recall that all computations at point  $x$  are performed conditionally on  $X'_{i_x}$  under **(CS2)**. Recall that we write  $\tilde{w} = w - X'_{i_x} + 2^{-j-1}$ . Given the independence of  $X'_{i_x}$  from the points of  $\mathcal{D}_n$ , we can write

$$\begin{aligned}
& \langle u, \mathbb{E}[Q_{\mathcal{H}_0} | X'_{i_x}] \cdot u \rangle \\
&= \int_{\mathbb{R}^d} \left( \sum_{\nu \in \mathfrak{S}} u_\nu \varphi_{j,\nu}(\tilde{w}) \mathbb{1}_{\mathcal{H}_0}(\tilde{w}) \right)^2 \mu(w) dw \\
&\geq \mu_{\min} \int_{\mathcal{A} \cap \mathcal{H}_0(X'_{i_x})} \left( \sum_{\nu \in \mathfrak{S}} u_\nu 2^{jd/2} \varphi(2^j(w - X'_{i_x}) + 2^{-1} - \nu) \right)^2 dw \\
&= \mu_{\min} \int_{2^j(\mathcal{A} - X'_{i_x}) \cap [-2^{-1}, 2^{-1}]^d} \left( \sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w + 2^{-1}) \right)^2 dw \\
&\geq \mu_{\min} \inf_{\mathfrak{m} \geq \mathfrak{m}_0} \inf_{z \in \mathcal{B}_\infty(0, \mathfrak{m})} \int_{\mathcal{B}_\infty(z, \mathfrak{m}) \cap [-2^{-1}, 2^{-1}]^d} \left( \sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w + 2^{-1}) \right)^2 dw \\
&= \mu_{\min} \inf_{\mathfrak{m} \geq \mathfrak{m}_0} \inf_{z \in \mathcal{B}_\infty(2^{-1}, \mathfrak{m})} \int_{\mathcal{B}_\infty(z, \mathfrak{m}) \cap [0, 1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_\nu \varphi_\nu(w) \right)^2 dw.
\end{aligned}$$

Thus, a direct application of [Lemma 12.1](#) leads to  $\lambda_{\min}(\mathbb{E}[Q_{\mathcal{H}_0} | X'_{i_x}]) \geq \mu_{\min} c_{\min}$ . This concludes the proof with  $g_{\min} = \mu_{\min} c_{\min}$ .  $\square$

*Proof of Proposition 12.5 under (CS2).* Recall first that  $\tilde{x} \in \mathcal{H}_0 = 2^{-j}[0, 1]^d$  and  $\nu(\mathcal{H}_0) = 0$ . At points  $x$  located at the boundary of  $\mathcal{A}$ , the cell  $\mathcal{H}_0(X'_{i_x}) = X'_{i_x} + 2^{-j}[-2^{-1}, 2^{-1}]^d$  centered in  $X'_{i_x}$  is eventually not a subset of  $\mathcal{A}$  anymore, so that the proof of Proposition 12.5 breaks down at the level of eq. (29). Using the independence of  $X'_{i_x}$  from the points of  $\mathcal{D}_n$ , we can write

$$\begin{aligned}
\mathbb{E}[Q_{\mathcal{H}_0}|X'_{i_x}] &= \int \varphi_{j,\nu(\mathcal{H}_0)}(\tilde{w})^2 \mu(w) dw \\
&\geq \mu_{\min} \int_{\mathcal{A}} 2^{jd} \varphi_{\nu(\mathcal{H}_0)}(2^j(w - X'_{i_x}) + 2^{-1})^2 dw \\
&= \mu_{\min} \int_{2^j(\mathcal{A} - X'_{i_x})} \varphi(w + 2^{-1})^2 dw \\
&\geq \mu_{\min} \inf_{\mathbf{m} \geq \mathbf{m}_0} \inf_{z \in \mathcal{B}_{\infty}(0, \mathbf{m})} \int_{\mathcal{B}_{\infty}(z, \mathbf{m})} \varphi(w + 2^{-1})^2 dw \\
&= \mu_{\min} \inf_{\mathbf{m} \geq \mathbf{m}_0} \inf_{z \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m})} \int_{\mathcal{B}_{\infty}(z, \mathbf{m})} \varphi(w)^2 dw \\
&\geq \mu_{\min} c_{\min} \geq 2t,
\end{aligned}$$

where the before last inequality comes as a direct application of Lemma 12.1, eq. (32). The end of the proof follows the same lines as previously, under setting (CS1).  $\square$

The updated proofs hinge on the following results.

**Lemma 12.1.** *Let  $r \in \mathbb{N}$ . Let  $\varphi$  be the Daubechies' scaling function of regularity  $r$  and  $\mathfrak{S} = \{\nu \in \mathbb{Z}^d : 2^{-1} \in \text{Supp} \varphi_{\nu}\}$ . Then, there exists a strictly positive absolute constant  $c_{\min}$  such that*

$$\inf_{u \in \mathbb{S}^{R^d-1}} \inf_{\mathbf{m} \geq \mathbf{m}_0} \inf_{z \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m})} \int_{\mathcal{B}_{\infty}(z, \mathbf{m}) \cap [0, 1]^d} \left( \sum_{\nu \in \mathfrak{S}} u_{\nu} \varphi_{\nu}(w) \right)^2 dw \geq c_{\min}, \quad (31)$$

$$\inf_{\mathbf{m} \geq \mathbf{m}_0} \inf_{z \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m})} \int_{\mathcal{B}_{\infty}(z, \mathbf{m})} \sum_{\nu \in \mathfrak{S}} \varphi_{\nu}(w)^2 dw \geq c_{\min}. \quad (32)$$

*Proof.* The proof of eq. (32) follows the same lines as the one of eq. (31). Therefore, we will only prove eq. (31). Now write  $u = (u_{\nu})_{\nu \in \mathfrak{S}}$  and  $\varphi_{\mathfrak{S}}(\cdot) = (\varphi_{\nu}(\cdot))_{\nu \in \mathfrak{S}}$ . Denote by  $\mathbb{S}^{R^d-1}$  the unit sphere of  $\mathbb{R}^{R^d}$ . Consider the map defined by

$$\begin{aligned}
f_{\mathbf{m}} : \mathcal{B}_{\infty}(2^{-1}, \mathbf{m}) \times \mathbb{S}^{R^d-1} &\mapsto \mathbb{R}, \\
f_{\mathbf{m}}(z, u) &= \int_{\mathcal{B}_{\infty}(z, \mathbf{m}) \cap [0, 1]^d} \langle u, \varphi_{\mathfrak{S}}(w) \rangle^2 dw.
\end{aligned}$$

Notice first that,  $u$  being fixed,  $\inf_{z \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m})} f_{\mathbf{m}}(z, u)$  is an increasing function of  $\mathbf{m}$ , since the integrand is positive and for  $\mathbf{m}_1 \geq \mathbf{m}_2$ , for all  $z_1 \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m}_1)$ , we can find  $z_2 \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m}_2)$  such that  $\mathcal{B}_{\infty}(z_2, \mathbf{m}_2) \subset \mathcal{B}_{\infty}(z_1, \mathbf{m}_1)$ . So that  $\inf_{\mathbf{m} \geq \mathbf{m}_0} \inf_{z \in \mathcal{B}_{\infty}(2^{-1}, \mathbf{m})} f_{\mathbf{m}}(z, u)$  reduces in fact to

$\inf_{z \in \mathcal{B}_\infty(2^{-1}, \mathbf{m}_0)} f_{\mathbf{m}_0}(z, u)$ . From now on, let us write  $f$  in place of  $f_{\mathbf{m}_0}$  and prove that  $f$  is continuous on its domain. We have indeed

$$|f(z_2, u_2) - f(z_1, u_1)| \leq |f(z_2, u_2) - f(z_1, u_2)| + |f(z_1, u_2) - f(z_1, u_1)|.$$

And

$$\begin{aligned} |f(z_1, u_2) - f(z_1, u_1)| &\leq \int_{\mathcal{B}_\infty(z_1, \mathbf{m}_0) \cap [0, 1]^d} |\langle u_2 - u_1, \varphi_\mathfrak{S}(w) \rangle| |\langle u_2 + u_1, \varphi_\mathfrak{S}(w) \rangle| dw \\ &\leq 2R^d \|\varphi\|_{\mathbb{L}_\infty(\mathbb{R}^d)}^2 \|u_2 - u_1\|_{\ell_2(\mathbb{R}^d)}. \end{aligned}$$

Besides, noticing that

$$\mathbb{1}_{\mathcal{B}_\infty(z_1, \mathbf{m}_0)}(w) = \mathbb{1}_{[0, 2\mathbf{m}_0]^d}(w - z_1 + \mathbf{m}_0) = \mathbb{1}_{[0, 1]^d}\left(\frac{w - z_1 + \mathbf{m}_0}{2\mathbf{m}_0}\right),$$

we can write

$$\begin{aligned} &|f(z_2, u_2) - f(z_1, u_2)| \\ &\leq \int_{[0, 1]^d} |\mathbb{1}_{\mathcal{B}_\infty(z_2, \mathbf{m}_0)}(w) - \mathbb{1}_{\mathcal{B}_\infty(z_1, \mathbf{m}_0)}(w)| |\langle u_2, \varphi_\mathfrak{S}(w) \rangle| dw \\ &\leq R^{d/2} \|\varphi\|_{\mathbb{L}_\infty(\mathbb{R}^d)} 2\mathbf{m}_0 \int_{\mathbb{R}^d} |\mathbb{1}_{[0, 1]^d}(w + \frac{\mathbf{m}_0 - z_2}{2\mathbf{m}_0}) - \mathbb{1}_{[0, 1]^d}(w + \frac{\mathbf{m}_0 - z_1}{2\mathbf{m}_0})| dw \\ &\leq 2R^{d/2} \|\varphi\|_{\mathbb{L}_\infty(\mathbb{R}^d)} \|z_2 - z_1\|_{\ell_1(\mathbb{R}^d)}, \end{aligned}$$

where the last inequality comes as a direct application of [Lemma 12.2](#). Since  $f$  is continuous and its domain is compact, it is in particular lower bounded and reaches its lower bound in  $(z^*, u^*) \in \mathcal{B}_\infty(2^{-1}, \mathbf{m}_0) \times \mathbb{S}^{r,d}$ . Now, since  $u^* \neq 0$ , it is a direct consequence of the local linear independence of the scaling functions  $(\varphi_{j,k})_{k \in \mathbb{Z}^d}$  described in [Proposition 12.8](#) that  $f(z^*, u^*) > 0$ . This concludes the proof.  $\square$

**Proposition 12.8.** *Let  $\mathbf{m}$  be a constant such that  $\mathbf{m} > 0$  and fix  $z \in \mathbb{R}^d$  such that  $z \in \mathcal{B}_\infty(2^{-1}, \mathbf{m})$ . Write  $\mathfrak{S}_d := \{k \in \mathbb{Z}^d : 2^{-1} \in \text{Supp} \varphi_k\}$ , the set of indexes corresponding to the scaling functions whose support  $\text{Supp} \varphi_k$  contains the point  $2^{-1} \in \mathbb{R}^d$ . The scaling functions  $(\varphi_k)$  verify the **local linear independence property** in the sense that  $\sum_{k \in \mathfrak{S}} \alpha_k \varphi_k = 0$  on the domain  $\mathcal{B}_\infty(z, \mathbf{m})$  if and only if  $\alpha_k = 0$  for all  $k \in \mathfrak{S}$ .*

*Proof.* The proof follows the same lines as the proof of Theorem 3.7.1 and Remark 3.7.1 in [\[10\]](#) and relies on arguments detailed in [\[36\]](#) and the references therein. For completeness, we detail the proof here.

In the one-dimensional case, the proof is given in [\[36, Corollaire 3\]](#). Notice indeed that for  $d = 1$ ,  $\lambda(\text{Supp} \varphi_k) = R := 2r - 1$  and  $\cup_{k \in \mathfrak{S}} \text{Supp} \varphi_k = [-R + 1, R]$ . Set  $\mathbb{1}_{[a, b]} = 0$  and  $[a, b] = \emptyset$  as soon as  $a > b$ . Now,  $\sum_{k \in \mathfrak{S}} \alpha_k \varphi_k = 0$  on  $[z - \mathbf{m}, z + \mathbf{m}]$  implies that both restrictions  $\mathbb{1}_{[-R+1, z-\mathbf{m}]} \sum_{k \in \mathfrak{S}} \alpha_k \varphi_k$  and  $\mathbb{1}_{[z+\mathbf{m}, R]} \sum_{k \in \mathfrak{S}} \alpha_k \varphi_k$  belong to  $\mathcal{V}_0$ , according to [\[36, Proposition 1\]](#). However  $z \in [1/2 - \mathbf{m}, 1/2 + \mathbf{m}]$ , therefore  $\lambda([-R + 1, z - \mathbf{m}]) < R$  and  $\lambda([z + \mathbf{m}, R]) < R$ , which



leads to a contradiction thanks to [36, Lemme 1, (ii)]. Therefore, both restrictions must be zero as well, which leads to  $\sum_{k \in \mathfrak{S}} \alpha_k \varphi_k = 0$  and thus  $\alpha_k = 0$  for all  $k \in \mathfrak{S}$ , since the  $(\varphi_k)_{k \in \mathbb{Z}}$  have disjoint supports.

The generalization to the  $d$ -dimensional case follows by induction as in the proof of Theorem 3.7.1 in [10]. Assume the property holds in dimension  $d - 1$  and let us work in dimension  $d$ . Let  $z \in \mathcal{B}_\infty(2^{-1}, \mathfrak{m})$  and, for any  $x \in \mathbb{R}^d$ , write  $x = (x_{1,d-1}, x_d)$ , where  $x_{1,d-1}$  is the element of  $\mathbb{R}^{d-1}$  consisting of the first  $d - 1$  coordinates of  $x$  and  $x_d$  is its last coordinate. Assume now that  $\sum_{k \in \mathfrak{S}_d} \alpha_k \varphi_k(x) = 0$  on  $\mathcal{B}_\infty(z, \mathfrak{m})$ . We can integrate this latter expression against a test function  $g(x_{1,d-1})$  supported on  $\mathcal{B}_\infty(z_{1,d-1}, \mathfrak{m})$  along the first  $d - 1$  dimensions. Using the fact that for all  $x \in \mathbb{R}^d$  and  $k \in \mathbb{Z}^d$ ,  $\varphi_k(x) = \varphi_{k_{1,d-1}}^{(d-1)}(x_{1,d-1}) \varphi_{k_d}^{(1)}(x_d)$ , we obtain

$$\begin{aligned} 0 &= \sum_{k_d \in \mathfrak{S}_1} a_{k_d} \varphi_{k_d}^{(1)}(x_d), \quad \text{on } \mathcal{B}_\infty(z_d, \mathfrak{m}), \\ a_{k_d} &= \sum_{m \in \mathfrak{S}_{d-1}} \int \alpha_{(m, k_d)} g(x_{1,d-1}) \varphi_m^{(d-1)}(x_{1,d-1}) dx_{1,d-1}. \end{aligned}$$

However, the linear independence property in one dimension triggers  $a_{k_d} = 0$  for all  $k_d \in \mathfrak{S}_1$ . Since the test function  $g$  is chosen arbitrarily, this latter result leads in turn to

$$\sum_{m \in \mathfrak{S}_{d-1}} \alpha_{(m, k_d)} \varphi_m^{(1, d-1)}(x_{1,d-1}) = 0, \quad \text{for all } k_d \in \mathfrak{S}_1 \text{ and } x_{1,d-1} \in \mathcal{B}_\infty(z_{1,d-1}, \mathfrak{m}).$$

Finally, using the induction hypothesis, we obtain  $\alpha_k = 0$  for all  $k \in \mathfrak{S}_d$ . □

**Lemma 12.2.** *Define the function  $f$  as follows*

$$\begin{aligned} f : \mathbb{R}^d &\mapsto \mathbb{R}^+, \\ f(y) &\triangleq \int_{\mathbb{R}^d} \mathbb{1}_{\{w+y \in [0,1]^d\}} dw. \end{aligned}$$

*Then,  $f$  is Lipschitz-continuous on  $\mathbb{R}^d$ , such that, for any two  $y, z \in \mathbb{R}^d$ ,*

$$|f(y) - f(z)| \leq 2 \|y - z\|_{\ell_1(\mathbb{R}^d)}.$$

*Proof.* The result is easily obtained by induction. We have indeed obviously for  $d = 1$ ,

$$\begin{aligned} |f(y) - f(z)| &\leq \int_{\mathbb{R}} |\mathbb{1}_{\{w+y \in [0,1]\}} - \mathbb{1}_{\{w+z \in [0,1]\}}| dw \\ &\leq \lambda([0, 1] - y) \Delta([0, 1] - z) \leq 2|y - z|. \end{aligned}$$

Now, write  $\mathfrak{T}(y) = [0, 1]^d - y$  and  $\mathfrak{T}_i(y) = [0, 1] - y_i$  for  $1 \leq i \leq d$ , where  $y_i$  stands for the  $i^{\text{th}}$  coordinate of  $y$ . In addition, for  $1 \leq p \leq q \leq d$ , we write

$$\mathfrak{T}_{p,q}(y) = \Pi_{i=p}^q \mathfrak{T}_i(y).$$

Recall that, for any  $y, z \in \mathbb{R}^d$ , we have the following relationships

$$\begin{aligned}\mathfrak{T}(y)\Delta\mathfrak{T}(z) &= (\mathfrak{T}(y) \cap \mathfrak{T}(z)^c) \cup (\mathfrak{T}(y)^c \cap \mathfrak{T}(z)), \\ \mathfrak{T}(y) &= (\mathfrak{T}_{1,d-1}(y) \times \mathbb{R}) \cap (\mathbb{R}^{d-1} \times \mathfrak{T}_d(y)), \\ (\mathfrak{T}_{1,d-1}(y) \times \mathbb{R})^c &= \mathfrak{T}_{1,d-1}(y)^c \times \mathbb{R}, \\ (\mathbb{R}^{d-1} \times \mathfrak{T}_d(y))^c &= \mathbb{R}^{d-1} \times \mathfrak{T}_d(y)^c.\end{aligned}$$

Making a repeated use of them, we can write

$$\begin{aligned}\mathfrak{T}(y)\Delta\mathfrak{T}(z) &= (\mathfrak{T}_{1,d-1}(y) \times \mathbb{R} \cap \mathbb{R}^{d-1} \times \mathfrak{T}_d(y))\Delta(\mathfrak{T}_{1,d-1}(z) \times \mathbb{R} \cap \mathbb{R}^{d-1} \times \mathfrak{T}_d(z)) \\ &= (\mathfrak{T}(y) \cap (\mathfrak{T}_{1,d-1}(z) \times \mathbb{R})^c) \cup (\mathfrak{T}(y) \cap (\mathbb{R}^{d-1} \times \mathfrak{T}_d(z))^c) \\ &\quad \cup (\mathfrak{T}(z) \cap (\mathfrak{T}_{1,d-1}(y) \times \mathbb{R})^c) \cup (\mathfrak{T}(z) \cap (\mathbb{R}^{d-1} \times \mathfrak{T}_d(y))^c) \\ &= (\mathfrak{T}(y) \cap (\mathfrak{T}_{1,d-1}(z)^c \times \mathbb{R})) \cup (\mathfrak{T}(y) \cap (\mathbb{R}^{d-1} \times \mathfrak{T}_d(z)^c)) \\ &\quad \cup (\mathfrak{T}(z) \cap (\mathfrak{T}_{1,d-1}(y)^c \times \mathbb{R})) \cup (\mathfrak{T}(z) \cap (\mathbb{R}^{d-1} \times \mathfrak{T}_d(y)^c)) \\ &= (\mathfrak{T}_{1,d-1}(y) \cap \mathfrak{T}_{1,d-1}(z)^c) \times \mathfrak{T}_d(y) \cup \mathfrak{T}_{1,d-1}(y) \times (\mathfrak{T}_d(y) \cap \mathfrak{T}_d(z)^c) \\ &\quad \cup (\mathfrak{T}_{1,d-1}(z) \cap \mathfrak{T}_{1,d-1}(y)^c) \times \mathfrak{T}_d(z) \cup \mathfrak{T}_{1,d-1}(z) \times (\mathfrak{T}_d(z) \cap \mathfrak{T}_d(y)^c).\end{aligned}$$

Now remark that each of the four above sets define hyper-rectangles of  $\mathbb{R}^d$ . Denote by  $\lambda^k$  the Lebesgue measure of  $\mathbb{R}^k$  and notice that  $\lambda^k(\mathfrak{T}_{1,k}(y)) = \lambda(\mathfrak{T}_k(y)) = 1$  for all  $1 \leq k \leq d$ . Therefore, we obtain

$$\begin{aligned}\lambda^d(\mathfrak{T}(y)\Delta\mathfrak{T}(z)) &\leq \lambda^{d-1}(\mathfrak{T}_{1,d-1}(y) \cap \mathfrak{T}_{1,d-1}(z)^c) + \lambda(\mathfrak{T}_d(y) \cap \mathfrak{T}_d(z)^c) \\ &\quad + \lambda^{d-1}(\mathfrak{T}_{1,d-1}(z) \cap \mathfrak{T}_{1,d-1}(y)^c) + \lambda(\mathfrak{T}_d(z) \cap \mathfrak{T}_d(y)^c) \\ &= \lambda^{d-1}(\mathfrak{T}_{1,d-1}(y)\Delta\mathfrak{T}_{1,d-1}(z)) + \lambda(\mathfrak{T}_d(y)\Delta\mathfrak{T}_d(z)).\end{aligned}$$

Consequently, we obtain immediately by induction that

$$|f(y) - f(z)| \leq \sum_{i=1}^d 2|y_i - z_i| = 2\|y - z\|_{\ell_1(\mathbb{R}^d)}.$$

□

### 12.3 Proof of the lower-bound

In this section, we show that the proof of [Theorem 7.3](#) under both settings **(CS1)** and **(CS2)** hinges on the well-known arguments of the proof of the lower-bound in the uniform design case. The arguments that follow are for the most part standard. Nonetheless, we report them here for completeness.

### 12.3.1 Preliminary results

First recall the definition of the Kullback-Leibler divergence for two probabilities  $\mathbb{P}$  and  $\mathbb{Q}$  on some measurable space  $(\Omega, \mathcal{F})$ :

$$\mathcal{K}(\mathbb{P}, \mathbb{Q}) = \begin{cases} \int_{\Omega} \log \left\{ \frac{d\mathbb{P}}{d\mathbb{Q}}(\omega) \right\} \mathbb{P}(d\omega) & \text{if } \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{otherwise} \end{cases}$$

Here we use a version of Fano's lemma that is due to [7]. It can be stated as follows

**Theorem 12.1** (Fano's Lemma). *Let  $\mathcal{F}$  be a  $\sigma$ -algebra on the space  $\Omega$ . Let  $G_i \in \mathcal{F}, i \in \{0, 1, \dots, m\}$  such that  $\forall i \neq j, G_i \cap G_j = \emptyset$ . Let  $\{\mathbb{P}_i\}_{i=0, \dots, m}$  be probability measures on  $(\Omega, \mathcal{F})$ . If we write*

$$\kappa(\mathbb{P}_0, \dots, \mathbb{P}_m) \triangleq \inf_{j=0, \dots, m} \frac{1}{m} \sum_{i \neq j} \mathcal{K}(\mathbb{P}_i, \mathbb{P}_j),$$

then, with  $C = e^{\frac{3}{e}}$ ,

$$\max_{i=0, \dots, m} \mathbb{P}_i(G_i^c) \geq \frac{1}{2} \wedge C \left\{ \sqrt{m} e^{-\kappa(\mathbb{P}_0, \dots, \mathbb{P}_m)} \right\}.$$

Furthermore, we will need the following two Propositions that give upper bounds on the Kullback-Leibler divergence under specific sets of assumptions.

**Proposition 12.9.** *Assume we work under (S2) and the Gaussian noise hypothesis (N2). Then, for any two  $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ -measurable maps  $\eta_1, \eta_2$ , we have*

$$\mathcal{K}(\mathbb{P}_{\eta_1}, \mathbb{P}_{\eta_2}) \leq \frac{1}{2} \|\eta_1 - \eta_2\|_{\mathbb{L}_2(\mathcal{E}, \mu)}^2.$$

*Proof.* The proof is straightforward and can be found in [9, Lemma 3.3]. □

**Proposition 12.10.** *Assume we work under any of the two settings (S1) and (S2) and the bounded noise assumption (N1). Assume moreover  $\xi$  to be independent of  $X$  and such that, for any Borel subset  $B$  of  $[-K, K]$ ,*

$$\mathbb{P}(\xi \in B) = C_K \int_B e^{-\frac{1}{2}u^2} du, \quad C_K^{-1} = \int_{-K}^K e^{-\frac{1}{2}u^2} du.$$

Then for any two  $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ -measurable maps  $\eta_1, \eta_2$ , we obtain

$$\mathcal{K}(\mathbb{P}_{\eta_1}, \mathbb{P}_{\eta_2}) \leq \max(4, 2K) C_K \|\eta_1 - \eta_2\|_{\mathbb{L}_2(\mathcal{E}, \mu)},$$

where  $\mathcal{B}(\mathcal{E})$  stands for the Borel  $\sigma$ -field of  $\mathcal{E}$ .

*Proof.* The proof is straightforward and hence reported to the Appendix. □

### 12.3.2 Lower bound under setting (CS1)

We choose  $\mu(\cdot) = 1$ , so that we are brought back to the case of a regression on a uniform design. Notice that, under (S1), such a  $\mu$  verifies (D1) since  $\mu_{\min}, \mu_{\max}$  must be such that  $\mu_{\min} \leq 1 \leq \mu_{\max}$ . In the case of a Gaussian noise such as under (N2), the proof of the lower-bound is well known and can be found in [42, 5, 9], to quote a few. In the case of a bounded noise, the same reasoning still applies thanks to Proposition 12.10 above.

### 12.3.3 Lower bound under setting (CS2)

In that case, a bit more work is needed. Let  $r \in \mathbb{N}$  and recall that  $2^{j_s} = \lfloor n^{\frac{1}{2s+d}} \rfloor$ . Let us write  $\mathcal{E} = \mathbb{R}^d$  and denote by  $(\mathcal{V}_j)_{j \geq 0}$  the Daubechies'  $r$ -MRA of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ . For any subset  $\mathcal{X}$  of  $\mathbb{Z}^d$ , write  $\nabla(\mathcal{X}) := [\{0, 1\} \setminus \{(0, \dots, 0)\}] \times \mathcal{X}$  and  $\nabla := \nabla(\mathbb{Z}^d)$ . We recall that the multivariate Daubechies' wavelets  $\psi_\nu, \nu \in \nabla$  associated with the  $r$ -MRA above are supported on hypercubes of volume  $R^d$ . We define the grid

$$G_{2^{j_s}} := \{(k_1, \dots, k_d) : k_i \in \{0, \dots, \lfloor \mathbf{m} 2^{j_s} \rfloor - 1\}\},$$

for some absolute constant  $\mathbf{m}$  to be chosen later. Obviously, we have  $\#G_{2^{j_s}} = \lfloor \mathbf{m} 2^{j_s} \rfloor^d$ . In addition, we take

$$\mathcal{A} = \bigcup_{k \in G_{2^{j_s}}} \mathcal{B}_\infty(k, \frac{1}{2} \lfloor \mathbf{m} 2^{j_s} \rfloor^{-1}).$$

Now, choose  $\mathbf{m}$  such that  $2^{j_s} 2^{-1} \lfloor \mathbf{m} 2^{j_s} \rfloor^{-1} \geq R$ , that is, say,  $\mathbf{m} = (2R)^{-1}$ . We assume that  $n$  is large enough, that is  $2^{j_s} \geq 2R$ , so that the grid is not degenerate. It is immediate that  $\mathcal{A}$  verifies (S2). Now, define  $\bar{\mu}$  on  $\mathbb{R}^d$  to be

$$\bar{\mu}(\cdot) = \sum_{k \in G_{2^{j_s}}} \mathbb{1}_{\mathcal{B}_\infty(k, \frac{1}{2} \lfloor \mathbf{m} 2^{j_s} \rfloor^{-1})}(\cdot).$$

By construction, we obtain  $\int_{\mathcal{A}} \bar{\mu}(w) dw = 1$ . And finally,  $\bar{\mu}$  verifies (D1) since  $\mu_{\min} \leq 1 \leq \mu_{\max}$ , as desired. Let  $j \geq j_s$ . Notice now that for any  $k \in G_{2^{j_s}}$ , we can find a subset  $G_{j,k}^*$  of  $\mathbb{Z}^d$  such that  $\#G_{j,k}^* = 2^{(j-j_s)d}$  and such that for all  $\nu \in \nabla(G_{j,k}^*)$ ,  $\text{Supp} \psi_{j,\nu} \subset \mathcal{B}_\infty(k, \frac{1}{2} \lfloor \mathbf{m} 2^{j_s} \rfloor^{-1})$ .

We write

$$\bar{G}_j := \bigcup_{k \in G_{2^{j_s}}} G_{j,k}^*.$$

In addition, we denote by  $\mathbf{a} := (a_\nu)_{\nu \in \nabla(\bar{G}_j)}$  an element of  $\{0, 1\}^{\nabla(\bar{G}_j)}$ . With these notations, we define

$$\Upsilon(\bar{G}_j) := \left\{ \eta \in \mathcal{L}^s(\mathbb{R}^d, M) : \eta = \gamma_j \sum_{\nu \in \nabla(\bar{G}_j)} a_\nu \psi_{j,\nu}, \mathbf{a} \in \{0, 1\}^{\bar{G}_j} \right\}, \quad j \geq j_s.$$

Now, we just need to remark that  $\#\nabla(\bar{G}_j) \approx 2^{jd}$  and for any two  $\eta_1, \eta_2 \in \Upsilon(\bar{G}_j)$ ,

$$\|\eta_1 - \eta_2\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})} = \|\eta_1 - \eta_2\|_{\mathbb{L}_p(\mathbb{R}^d, \lambda)}. \quad (33)$$

This last result together with both [Proposition 12.9](#) and [Proposition 12.10](#) above allows us to conclude once more using the well known arguments of the proof of the lower bound in a uniform design, under both assumptions **(N1)** and **(N2)**. For completeness, we detail the end of the reasoning below. We denote by **(Ind)** the assumption that  $\xi$  is independent of  $X$  and write

$$\mathcal{P}^*(\bar{G}_j) := \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}) : \eta \in \Upsilon(\bar{G}_j), \mu = \bar{\mu}, \text{(\textbf{Ind}) holds true}\}.$$

We denote by  $\eta_{\mathbf{a}}$  the elements of  $\Upsilon(\bar{G}_j)$ , where we write  $\mathbf{a} \in \{0, 1\}^{\nabla(\bar{G}_j)}$ . It appears that, for  $\eta_{\mathbf{a}} \in \Upsilon(\bar{G}_j)$ ,

$$\|\eta_{\mathbf{a}}\|_{\mathcal{L}^s(\mathbb{R}^d)} \leq C|\gamma_j|2^{j(s+\frac{d}{2})} \sup_{\nu \in \nabla(\bar{G}_j)} a_\nu \leq C|\gamma_j|2^{j(s+\frac{d}{2})},$$

where the first inequality is a direct application of [eq. \(37\)](#). Thus, in order to have  $\eta_{\mathbf{a}} \in \mathcal{L}^s(\mathbb{R}^d, M)$ , it is enough to choose  $|\gamma_j| \leq CM2^{-j(s+\frac{d}{2})}$ . With this constraint on  $\gamma_j$ , we obtain

$$\mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^r) \supset \mathcal{P}^*(\bar{G}_j).$$

Besides, for any two  $\mathbf{a}, \mathbf{a}' \in \{0, 1\}^{\nabla(\bar{G}_j)}$ , a direct application of [[10](#), eq. (3.6.16)] leads to

$$\|\eta_{\mathbf{a}} - \eta_{\mathbf{a}'}\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})} \geq c|\gamma_j|2^{jd(\frac{1}{2}-\frac{1}{p})} \left( \sum_{\nu \in \nabla(\bar{G}_j)} |a_\nu - a'_\nu|^p \right)^{\frac{1}{p}}. \quad (34)$$

The Varshamov-Gilbert Lemma (see [[51](#), Lemma 2.9]) ensures there exists a subset  $\nabla(A_j) \subset \nabla(\bar{G}_j)$  such that  $\#\nabla(A_j) \geq \frac{\#\nabla(\bar{G}_j)}{8}$  and  $\sum_{\nu \in \nabla(A_j)} |a_\nu - a'_\nu| \geq \frac{\#\nabla(\bar{G}_j)}{8}$ . Write  $\Gamma_j := \{(a_\nu)_{\nu \in \nabla(\bar{G}_j)} : a_\nu \in \{0, \mathbb{1}_{\{\nu \in \nabla(A_j)\}}\}\}$ . Now, if we pick  $|\gamma| = c2^{-j(s+\frac{d}{2})}$  and  $\mathbf{a}, \mathbf{a}' \in \Gamma_j$ , [eq. \(34\)](#) together with the latter results leads to

$$\|\eta_{\mathbf{a}} - \eta_{\mathbf{a}'}\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})} \geq c|\gamma|2^{jd(\frac{1}{2}-\frac{1}{p})} \left( \frac{2^{jd}}{8} \right)^{\frac{1}{p}} = c2^{-js}. \quad (35)$$

Now we can write  $\delta := \frac{1}{2}c2^{-js}$  and  $G_{\mathbf{a}} = \{\|\theta - \eta_{\mathbf{a}}\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})} < \delta\}$ . It results from [eq. \(35\)](#) that, for  $\mathbf{a} \in \Gamma_j$ , the  $G_{\mathbf{a}}$ 's are mutually exclusive. Let us denote by  $\mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}$  the law of  $(X, Y)$  under the assumption that  $\eta = \eta_{\mathbf{a}}$  and  $\mu = \bar{\mu}$  and  $\mathbb{E}_{\eta_{\mathbf{a}}, \bar{\mu}}$  the corresponding expectation. We will add the superscript  $\otimes n$  when dealing with their  $n^{th}$  tensorial product. Thus we can apply [Theorem 12.1](#) to obtain

$$\begin{aligned} \max_{\mathbb{P} \in \mathcal{P}^*(\Upsilon(A_j))} \mathbb{E} \|\theta - \eta\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})}^p &= \max_{\mathbf{a} \in \Gamma_j} \mathbb{E}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n} \|\theta - \eta_{\mathbf{a}}\|_{\mathbb{L}_p(\mathcal{A}, \bar{\mu})}^p \\ &\geq \delta^p \max_{\mathbf{a} \in \Gamma_j} \mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n} (\|\theta - \eta_{\mathbf{a}}\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})} \geq \delta) \end{aligned} \quad (36)$$

$$\begin{aligned}
&= \delta^p \max_{\mathbf{a} \in \Gamma_j} \mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n}(G_{\mathbf{a}}^c) \\
&\geq \delta^p \left\{ \frac{1}{2} \wedge C \sqrt{2^{\#\Gamma_j} - 1} e^{-\kappa((\mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n})_{\mathbf{a} \in \Gamma_j})} \right\}.
\end{aligned}$$

In addition both [Proposition 12.9](#) and [Proposition 12.10](#) lead to

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n}, \mathbb{P}_{\eta_{\mathbf{a}'}, \bar{\mu}}^{\otimes n}) &= n \mathcal{K}(\mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}, \mathbb{P}_{\eta_{\mathbf{a}'}, \bar{\mu}}) \leq Cn \|\eta_{\mathbf{a}} - \eta_{\mathbf{a}'}\|_{\mathbb{L}_2(\mathcal{E}, \bar{\mu})}^2 \\
&= Cn |\gamma_j|^2 \left\| \sum_{\nu \in \nabla(A_j)} (a_\nu - a'_\nu) \psi_{j, \nu} \right\|_{\mathbb{L}_2(\mathcal{E}, \lambda)}^2 \\
&\leq Cn |\gamma_j|^2 \sum_{\nu \in \nabla(A_j)} |a_\nu - a'_\nu|^2 \leq Cn |\gamma_j|^2 2^{jd} = Cn 2^{-2js},
\end{aligned}$$

where the second equality makes use of [eq. \(33\)](#). As a consequence

$$\sqrt{2^{\#\Gamma_j} - 1} e^{-\kappa((\mathbb{P}_{\eta_{\mathbf{a}}, \bar{\mu}}^{\otimes n})_{\mathbf{a} \in \Gamma_j})} \geq 2^{c2^{jd}} e^{-cn2^{-2js}}.$$

In order to balance this last term, it is enough to choose  $j$  such that  $n2^{-2js} = c2^{jd}$ , which leads to  $n^{-\frac{1}{2}} = c2^{-j(s+\frac{d}{2})}$ . Combining this value with [eq. \(36\)](#) and the inclusion  $\mathcal{L}^s(\mathbb{R}^d, M) \supset \Upsilon(\bar{G}_j) \supset \Upsilon(A_j)$  gives

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathbf{CS2}, \mathbf{H}_s^{\mathbf{r}})} \mathbb{E} \|\theta - \eta\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})}^p \geq \max_{\mathbb{P} \in \mathcal{P}^*(\Upsilon(A_j))} \mathbb{E} \|\theta - \eta\|_{\mathbb{L}_p(\mathcal{E}, \bar{\mu})}^p \geq c(2^{-j})^{sp} = cn^{-\frac{sp}{2s+d}}.$$

This concludes the proof of [Theorem 7.3](#). □

## 13 Appendix

### 13.1 Generalized Lipschitz spaces

Here, we sum up relevant facts about Lipschitz and Besov spaces on  $\mathbb{R}^d$  as stated in [\[10, Chap. 3\]](#) for any  $d \in \mathbb{N}$  and [\[14, Chap. 2, §9\]](#) for  $d = 1$ . Let us denote by  $\mathcal{C}(\mathbb{R}^d)$  and  $\tilde{\mathcal{C}}(\mathbb{R}^d)$  the spaces of continuous and absolutely continuous functions on  $\mathbb{R}^d$ , respectively. Let us denote by  $\|\cdot\|$  the Euclidean norm of  $\mathbb{R}^d$ ,  $f$  a function defined on  $\mathbb{R}^d$  and write  $\Delta_h^1(f, x) = |f(x+h) - f(x)|$  for any  $x \in \mathbb{R}^d$ . For any  $r \in \mathbb{N}$  and all  $x \in \mathbb{R}^d$ , we further define the  $r^{th}$ -finite difference by induction as follows,

$$\Delta_h^r(f, x) = \Delta_h^1(\Delta_h^{r-1}(f, x)),$$

and the  $r^{th}$ -modulus of smoothness of  $f \in \mathcal{C}(\mathbb{R}^d)$  as follows

$$\omega_r(f, t)_\infty = \sup_{0 \leq \|h\| \leq t} \|\Delta_h^r(f, \cdot)\|_{\mathbb{L}_\infty(\mathbb{R}^d, \lambda)}.$$

Write  $s > 0$  and  $r = \lfloor s \rfloor + 1$ . The Besov space  $B_{\infty,\infty}^s$  on  $\mathbb{R}^d$ , also known as the generalized Lipschitz space  $\mathcal{L}^s(\mathbb{R}^d)$ , is the collection of all functions  $f \in \tilde{\mathcal{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d, \lambda)$  such that the semi-norm

$$|f|_{\mathcal{L}^s(\mathbb{R}^d)} := \sup_{t>0} (t^{-s} \omega_r(f, t)_\infty),$$

is finite. The norm for  $\mathcal{L}^s(\mathbb{R}^d)$  is subsequently defined as

$$\|f\|_{\mathcal{L}^s(\mathbb{R}^d)} := \|f\|_{\mathbb{L}_\infty(\mathbb{R}^d, \lambda)} + |f|_{\mathcal{L}^s(\mathbb{R}^d)}.$$

Fix a real number  $M > 0$ . Throughout the paper,  $\mathcal{L}^s(\mathbb{R}^d, M)$  refers to the ball of  $\mathcal{L}^s(\mathbb{R}^d)$  of radius  $M$ . Obviously, the elements of  $\mathcal{L}^s(\mathbb{R}^d, M)$  are  $\lambda$ -a.e. uniformly bounded by  $M$  on  $\mathbb{R}^d$ . As described in [14, 10], there exists an alternative definition of Lipschitz spaces  $\mathcal{C}^s(\mathbb{R}^d)$ , also known as Hölder spaces, which goes as follows. For any integer  $d$ , multi-index  $q = (q_1, \dots, q_d) \in \mathbb{N}^d$  and  $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ , we define the differential operator  $\partial^q$  as usual by  $\partial^q := \frac{\partial^{q_1+\dots+q_d}}{\partial^{q_1} x_1 \dots \partial^{q_d} x_d}$ . For any positive integer  $s$ ,  $\mathcal{C}^s(\mathbb{R}^d)$  consists of the functions  $f$  on  $\mathbb{R}^d$  such that  $\partial^q f$  is bounded and absolutely continuous on  $\mathbb{R}^d$ , for all  $q \in \mathbb{N}^d$  such that  $|q|_1 := q_1 + \dots + q_d \leq s$ . This definition is extended to non-integer  $s$  as follows,

$$\begin{aligned} \mathcal{C}^s(\mathbb{R}^d) &:= \{f \in \tilde{\mathcal{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d, \lambda) : \sup_{x \in \mathbb{R}^d} \Delta_h^1(f, x) \leq C|h|^s\}, \quad 0 < s < 1, \\ \mathcal{C}^s(\mathbb{R}^d) &:= \{f \in \tilde{\mathcal{C}}(\mathbb{R}^d) \cap \mathbb{L}_\infty(\mathbb{R}^d, \lambda) : \\ &\quad \partial^q f \in \mathcal{C}^{s-m}(\mathbb{R}^d), |q|_1 = m\}, \quad m < s < m+1, \quad m \in \mathbb{N}. \end{aligned}$$

It can be shown that, for all non-integer  $s > 0$ ,  $\mathcal{C}^s(\mathbb{R}^d) = \mathcal{L}^s(\mathbb{R}^d)$ , while  $\mathcal{C}^s(\mathbb{R}^d)$  is a strict subset of  $\mathcal{L}^s(\mathbb{R}^d)$  when  $s \in \mathbb{N}$  (see [14, p. 52] for examples of functions that belong to  $\mathcal{L}^1([0, 1])$  but not to  $\mathcal{C}^1([0, 1])$  in the particular case where  $d = 1$ ).

Furthermore, we define these function spaces on the subset  $\mathcal{E}$  of  $\mathbb{R}^d$  as the restriction of their elements to  $\mathcal{E}$ . As explained in [10, Remark 3.2.4], function spaces on  $\mathcal{E}$  can be defined by restriction or, alternatively, in an intrinsic way, and both definitions coincide for fairly general domains  $\mathcal{E}$  of  $\mathbb{R}^d$ .

Looking at function spaces on  $\mathcal{E}$  as function spaces on  $\mathbb{R}^d$  restricted to  $\mathcal{E}$  justifies the use of MRAs of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  in our local analysis.

## 13.2 MRAs and smoothness analysis

### 13.2.1 Generalities

In order to estimate  $\eta$  on  $\mathcal{E}$ , we will work with MRAs of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$  built upon a compactly supported scaling function  $\varphi$ . Multivariate MRAs will always be assumed to be obtained from a tensorial product of one-dimensional MRAs, as described in [10, §1.4, eq. (1.4.10)] and recalled below. We will denote by  $\varphi_{j,k}(\cdot) = 2^{jd/2} \varphi(2^j \cdot - k)$  the translated and dilated version of  $\varphi$  with  $k \in \mathbb{Z}^d$ . As usual, we denote  $\mathcal{V}_j = \text{Closure}(\text{Span}\{\varphi_{j,k}, k \in \mathbb{Z}^d\})$ , so that  $\text{Closure}(\cup_{j \geq 0} \mathcal{V}_j) =$

$\mathbb{L}_2(\mathbb{R}^d, \lambda)$  (where the closures are taken with respect to the  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ -metric). As we are interested in finer and finer approximations of an objective function as the sample size grows, we can in fact truncate the MRA at resolution scales  $2^{-j}$ ,  $j \geq 0$ . In the sequel, we will still refer to it as a MRA.

Here, let us simply recall that, for any  $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$  and  $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ , the multivariate scaling functions write as

$$\varphi_{j,k}(x) = \varphi_{j,k_1}(x_1) \dots \varphi_{j,k_d}(x_d).$$

Let us denote by  $\psi$  the univariate wavelet associated to the univariate scaling function  $\varphi$ . Then the multivariate wavelets  $\psi_{j,\nu}$  are indexed by  $j \geq 0$  and the double index  $\nu = (\varepsilon, k) \in \nabla = \{0, 1\}^d \setminus \{(0, \dots, 0)\} \times \mathbb{Z}^d$  and write as

$$\psi_{j,\nu}(x) = \psi_{j,k}^{(\varepsilon_1)}(x_1) \dots \psi_{j,k}^{(\varepsilon_d)}(x_d),$$

where we have written  $\psi^{(0)} = \varphi$ ,  $\psi^{(1)} = \psi$  and  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$ .

### 13.2.2 Connection between MRAs and generalized Lipschitz spaces

The  $r$ -MRAs defined in [Section 5.2](#) are intimately connected with generalized Lipschitz spaces. Assume we are given a  $r$ -MRA with  $r \in \mathbb{N}$  and  $\eta \in \mathcal{L}^s(\mathbb{R}^d, M)$ , where  $s \in (0, r)$  and  $M > 0$ . Denote by  $\mathcal{P}_j\eta$  the projection of  $\eta$  onto  $\mathcal{V}_j$  and by  $\mathcal{R}_j\eta = \eta - \mathcal{P}_j\eta$  the corresponding remainder. Then, we have for all  $x \in \mathbb{R}^d$ ,  $\eta(x) = \mathcal{P}_j\eta(x) + \mathcal{R}_j\eta(x)$  where  $\|\mathcal{R}_j\eta\|_{\mathbb{L}_\infty(\mathbb{R}^d, \lambda)} \leq M2^{-js}$ , as detailed in [\[10, Corollary 3.3.1\]](#). It is noteworthy that the above approximation results remain valid in the particular case where we work on the subset  $\mathcal{E}$  of  $\mathbb{R}^d$  and consider  $\eta$  to be the restriction to  $\mathcal{E}$  of an element of  $\mathcal{L}^s(\mathbb{R}^d)$ .

Let us denote by  $\{\varphi_k : k \in \mathbb{Z}^d\} \cup \{\psi_{j,\nu} : j \geq 0, \nu \in \nabla\}$  the Daubechies' wavelet basis associated to a  $r$ -MRA (see [Section 13.2](#) for notations and [\[11, Chap. 5, 6\]](#) for a detailed review of wavelet bases and [\[10, §1.4, Remark 1.4.5\]](#) for the construction of their multi-dimensional isotropic version). For any  $\eta \in \mathcal{L}^s(\mathbb{R}^d)$ , we write  $\alpha_k := \langle \eta, \varphi_k \rangle_{\mathbb{L}_2(\mathbb{R}^d, \lambda)}$  and  $\beta_{j,\nu} := \langle \eta, \psi_{j,\nu} \rangle_{\mathbb{L}_2(\mathbb{R}^d, \lambda)}$ . Then, we have the following well-known norm equivalence

$$\|\eta\|_{\mathcal{L}^s(\mathbb{R}^d)} \approx \sup_{k \in \mathbb{Z}^d} |\alpha_k| + \sup_{j \geq 0} 2^{j(s+\frac{d}{2})} \sup_{\nu \in \nabla} |\beta_{j,\nu}|, \quad s \in (0, r). \quad (37)$$

This latter result happens to be useful in the proof of the lower bound (see [Section 12.3](#)).

## 13.3 A technical comment relative to the lattice structure of the MRA under (CS1)

Assume we work under (CS1) with the Daubechies'  $r$ -MRA of  $\mathbb{L}_2(\mathbb{R}^d, \lambda)$ . Let  $x \in [0, 1]^d$ . Notice that the scaling functions  $\varphi_{j,k}$ ,  $k \in \mathbb{Z}^d$  are supported on hypercubes of the form  $\text{Supp}\varphi_{j,k} = 2^{-j}(N_{1,k}, N_{2,k})$ , where  $N_{1,k}, N_{2,k} \in \mathbb{Z}^d$ . When doing estimation at point  $x$  at resolution  $2^{-j}$ , we have to differentiate two cases according to whether  $x$  belongs to  $\text{Inter}(\text{Supp}\varphi_{j,k})$  or  $\partial\text{Supp}\varphi_{j,k}$ .



This is due to the fact that  $\#\mathcal{S}_j(x)$  is constant on each open hypercube of  $2^{-j}[\mathbb{Z}^d + (0, 1)^d]$  while it raises ( $r = 1$ ) or drops ( $r \geq 2$ ) at their boundary. However, notice that it is only necessary to focus on the case where  $x \in \text{Inter}(\text{Supp}\varphi_{j,k})$ , since we are interested in results that hold  $\mathbb{P}_X$ -almost surely.  $\mathbb{P}_X$  is indeed absolutely continuous with respect to the Lebesgue measure  $\lambda$  on  $\mathbb{R}^d$  (with density  $\mu$ ) and for any  $j \geq 0$ ,  $k \in \mathbb{Z}^d$ ,  $\lambda(\partial\text{Supp}\varphi_{j,k}) = 0$ . Consequently

$$\mathbb{P}_X \left( \bigcup_{j \geq 0, k \in \mathbb{Z}^d} \partial\text{Supp}\varphi_{j,k} \right) = 0.$$

When working under (S1), we implicitly assume throughout the paper that

$$\mathcal{A} = [0, 1]^d \setminus \bigcup_{j \geq 0, k \in \mathbb{Z}^d} \partial\text{Supp}\varphi_{j,k}.$$

Notice that this problem does not occur under (CS2) where, by construction,  $x$  is always located inside  $\mathcal{H}_0(X_{i_x})$ .

### 13.4 Proof of Proposition 12.10

*Proof of Proposition 12.10.* For any two Borel subsets  $A$  and  $B$  of  $\mathbb{R}$  and  $\mathcal{E}$ , respectively, and any  $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ -measurable map  $\eta$ , a straightforward computation leads to

$$\mathbb{P}_\eta(Y \in A, X \in B) = \int_B \mu(x) dx \int_A du C_K e^{-\frac{1}{2}(u - \eta(x))^2}.$$

Therefore, for any two  $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ -measurable maps  $\eta_1, \eta_2$ , one can write

$$\begin{aligned} & \mathcal{K}(\mathbb{P}_{\eta_1}, \mathbb{P}_{\eta_2}) \\ &= \int_{\Omega} \log \left( \frac{d\mathbb{P}_{\eta_1}}{d\mathbb{P}_{\eta_2}}(\omega) \right) d\mathbb{P}_{\eta_1}(\omega) \\ &= \frac{1}{2} \int_{\mathcal{A}} dx \mu(x) \int_{-K}^K du C_K [(u - \eta_2(x))^2 - (u - \eta_1(x))^2] e^{-\frac{1}{2}(u - \eta_1(x))^2} \\ &= \frac{1}{2} \int_{\mathcal{A}} dx \mu(x) (\eta_1(x) - \eta_2(x))^2 \int_{-K}^K du C_K e^{-\frac{1}{2}(u - \eta_1(x))^2} \\ &+ \int_{\mathcal{A}} dx \mu(x) (\eta_1(x) - \eta_2(x)) \int_{-K}^K du C_K (u - \eta_1(x)) e^{-\frac{1}{2}(u - \eta_1(x))^2}. \end{aligned}$$

Now, notice first that

$$\int_{-K}^K du C_K e^{-\frac{1}{2}(u - \eta_1(x))^2} \leq 2K C_K.$$

In addition, notice that

$$\left| \int_{-K}^K du C_K(u - \eta_1(x)) e^{-\frac{1}{2}(u - \eta_1(x))^2} \right| = C_K |e^{-\frac{1}{2}(K - \eta_1(x))^2} - e^{-\frac{1}{2}(K - \eta_1(x))^2}| \leq 2C_K.$$

Consequently, we end up with the desired result

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{\eta_1}, \mathbb{P}_{\eta_2}) &= 2C_K \|\eta_1 - \eta_2\|_{\mathbb{L}_1(\mathcal{E}, \mu)} + KC_K \|\eta_1 - \eta_2\|_{\mathbb{L}_2(\mathcal{E}, \mu)} \\ &\leq \max(4, 2K)C_K \|\eta_1 - \eta_2\|_{\mathbb{L}_2(\mathcal{E}, \mu)}. \end{aligned}$$

□

### 13.5 Proof of the classification results

*Proof of Proposition 10.1.* Notice first that eq. (18) is a direct consequence of Theorem 9.1. The proof of eq. (19) follows the same lines as the proof of Theorem 7.2. The reader is therefore advised to read the latter proof before this one. Notice that

$$\begin{aligned} \mathbb{P}(|\eta_{j_s^\star(x)}^\star(x) - \eta(x)| \geq \delta) &\leq \mathbb{P}(|\eta_{j_s}^\star(x) - \eta(x)| \geq \delta/3) \\ &\quad + \mathbb{P}(|\eta_{j_s^\star(x)}^\star(x) - \eta_{j_s}^\star(x)| \geq \delta/3, j_s^\star(x) \leq j_s) \\ &\quad + \mathbb{P}(|\eta_{j_s^\star(x)}^\star(x) - \eta_{j_s}^\star(x)| \geq \delta/3, j_s^\star(x) > j_s). \end{aligned}$$

Let us denote the three terms on the rhs from top to bottom by *I*, *II* and *III*. *I* is upper bounded using eq. (18) above with  $j = j_s$  and  $\delta > 6 \cdot 2^{-j_s s} \max(1, 3\pi_0 R^d \mu_{\max})$ . This leads to

$$I \leq 3R^{2d} n^{-\kappa C'_{10}} + 2R^d \exp(-C_{10} a_n(j_s) \delta^2),$$

where  $C'_{10}$  and  $C_{10}$  are such that

$$C'_{10} := \frac{1}{4\mu_{\max} R^{4d} + \frac{8}{3} R^{2d} \pi_0^{-1}}, \quad C_{10} := \frac{C_6(r, d, \mu_{\max}, 1, \pi_0)}{2\pi_0^2}, \quad (38)$$

and  $C_6$  is defined in eq. (25). Now, recall from the proof of Theorem 7.2 that, on the event  $\{j_s^\star \leq j_s\}$ , *II* = 0 as soon as  $\delta > 6t(n)2^{j_s \frac{d}{2}} n^{-\frac{1}{2}}$ . Let us finally turn to *III*. With the same notations as in the proof of Theorem 7.2, we obtain

$$\begin{aligned} III &\leq \sum_{j \in \mathcal{J}_n(j_s)} \sum_{k \in \mathcal{J}_n(j-1)} \mathbb{P}(\mathcal{G}(x, j-1, k)) \\ &\leq (\log n)^2 10R^{2d} n^{-\kappa C_6}, \end{aligned}$$

where the last inequality uses Proposition 12.2 and the fact that  $\#\mathcal{J}_n \leq \log n$ . Denote moreover,

$$C'_6 = \frac{1}{64\mu_{\max} R^{4d} + 8R^{2d} \pi_0^{-1}}, \quad C_{11} := 2^{-2r} C_{10}, \quad (39)$$

and notice finally that  $C'_{10} \wedge C_6 > C'_6$ . Besides, recall from eq. (10c) that  $2^{-j_s s} \leq 2^r n^{-\frac{s}{2s+d}}$ , so that for  $n$  large enough,  $6 \cdot 2^{-j_s s} \max(1, 3\pi_0 R^d \mu_{\max}) \leq 62^r t(n) n^{-\frac{s}{2s+d}}$  and a direct application of eq. (10c) leads to  $C_{10} a_n(j_s) \geq C_{11} n^{\frac{2s}{2s+d}}$ . This concludes the proof. □

*Proof of Theorem 10.1.* It follows closely the proof of [4, Theorem 3.1]. Consider the sets  $A_k$  defined as

$$\begin{aligned} A_0 &:= \{\tau \in \mathcal{E} : 0 < |\eta(x) - \frac{1}{2}| \leq \delta_n\}, \\ A_k &:= \{\tau \in \mathcal{E} : 2^{k-1}\delta_n < |\eta(x) - \frac{1}{2}| < 2^k\delta_n\}, \quad 1 \leq k \leq \bar{k}, \end{aligned}$$

where we define  $\bar{k}$  by  $2^{\bar{k}+1} = \lceil \delta_n^{-1} \rceil$ . Notice indeed that  $2^{k-1}\delta_n \geq 2^{-1}$ , that is  $A_k = \emptyset$  for  $k \geq \bar{k} + 1$ . Therefore, we can write

$$\begin{aligned} \mathcal{R}(h_n) &= \sum_{0 \leq k \leq \bar{k}} \mathbb{E}(|2\eta(X) - 1| \mathbb{1}_{\{h_n(X) \neq h^*(X)\}} \mathbb{1}_{\{X \in A_k\}}) \\ &\leq 2\delta_n \mathbb{P}(0 < |\eta(X) - \frac{1}{2}| \leq \delta_n) \\ &\quad + \sum_{1 \leq k \leq \bar{k}} \mathbb{E}(|2\eta(X) - 1| \mathbb{1}_{\{h_n(X) \neq h^*(X)\}} \mathbb{1}_{\{X \in A_k\}}). \end{aligned}$$

Notice now that we have  $|\eta - \frac{1}{2}| \leq |\eta_n - \eta|$  on the event  $\{h_n \neq h^*\}$ . So that, for any  $1 \leq k \leq \bar{k}$ , we obtain

$$\begin{aligned} &\mathbb{E}(|2\eta(X) - 1| \mathbb{1}_{\{h_n(X) \neq h^*(X)\}} \mathbb{1}_{\{X \in A_k\}}) \\ &\leq 2^{k+1}\delta_n \mathbb{E}(\mathbb{1}_{\{|\eta_n(X) - \eta(X)| \geq 2^{k-1}\delta_n\}} \mathbb{1}_{\{0 < |\eta(X) - \frac{1}{2}| \leq 2^k\delta_n\}}) \\ &= 2^{k+1}\delta_n \mathbb{E}\mathbb{P}^{\otimes n}(|\eta_n(X) - \eta(X)| \geq 2^{k-1}\delta_n | X) \mathbb{1}_{\{0 < |\eta(X) - \frac{1}{2}| \leq 2^k\delta_n\}} \\ &\leq 2^{k+1}\delta_n \mathbb{P}(0 < |\eta(X) - \frac{1}{2}| \leq 2^k\delta_n) (C_{11}b_n + C_{12} \exp(-C_{13}a_n(2^{k-1}\delta_n)^2)) \\ &\leq 2C_*(2^k\delta_n)^{1+\vartheta} (C_{11}b_n + C_{12} \exp(-C_{13}a_n(2^{k-1}\delta_n)^2)), \end{aligned}$$

where the before last inequality uses [eq. \(20\)](#), which is possible since  $2^{k-1}\delta_n \leq 1$  as long as  $k \leq \bar{k}$ , and the last one uses **(MA)**. So that putting everything together, we obtain

$$\begin{aligned} \mathcal{R}(h_n) &\leq 2C_*\delta_n^{1+\vartheta} + 2C_*C_{11}(2^{\bar{k}}\delta_n)^{1+\vartheta}b_n \\ &\quad + 2C_*C_{12}\delta_n^{1+\vartheta} \sum_{1 \leq k \leq \bar{k}} 2^{k(1+\vartheta)} \exp(-C_{13}a_n\delta_n^2 2^{2(k-1)}). \end{aligned}$$

It remains to notice that  $2^{\bar{k}}\delta_n \leq 1$  to conclude the proof. □

*Proof of Corollary 10.1.* The proof of [eq. \(21\)](#) and [eq. \(22\)](#) are straightforward consequences of [Proposition 10.1](#) and [Theorem 10.1](#). Start with [eq. \(21\)](#) and choose  $\delta_n = 32^{-j_{ss}} \max(1, 3\pi_0 R^d \mu_{\max})$ . Notice first that [eq. \(10a\)](#) leads to

$$a_n(j_s)\delta_n^2 \geq 9 \max(1, 3\pi_0 R^d \mu_{\max})^2$$

Besides, [eq. \(10c\)](#) leads to  $2^{-j_s s} \leq 2^r n^{-\frac{s}{2s+d}}$ . Finally, notice that

$$\delta_n^{-(1+\vartheta)} n^{-\kappa C'_{10}} \leq C n^{\frac{s}{2s+d}(1+\vartheta) - \kappa C'_{10}}$$

which goes to zero as soon as, say,  $\kappa > \frac{1+\vartheta}{2C'_{10}}$ . This concludes the proof of [eq. \(21\)](#) with

$$C_{15} := 2C_* C_{14}^{1+\vartheta} 2^{r(1+\vartheta)} \left( 2 + 2R^d \sum_{k \geq 1} 2^{k(1+\vartheta)} \exp(-C_{10} C_{14}^2 2^{2(k-1)}) \right), \quad (40)$$

$$C_{14} := 3 \max(1, 3\pi_0 R^d \mu_{\max}), \quad (41)$$

where  $C_{10}$  is defined in [eq. \(38\)](#) above. Let us now turn to [eq. \(22\)](#) and choose  $\delta_n = t(n) n^{-\frac{s}{2s+d}}$ . Notice now that

$$n^{\frac{2s}{2s+d}} \delta_n^2 \geq t(n)^2$$

and  $2R^d \sum_{k \geq 1} 2^{k(1+\vartheta)} \exp(-C_{11} t(n)^2 2^{2(k-1)})$  can be made as small as we want for  $n$  large enough. Finally, notice that

$$(\log n)^2 \delta_n^{-(1+\vartheta)} n^{-\kappa C'_6} \leq C (\log n)^2 t(n)^{-(1+\vartheta)} n^{\frac{s}{2s+d}(1+\vartheta) - \kappa C'_6}$$

which goes to zero as soon as, say,  $\kappa > \frac{1+\vartheta}{2C'_6}$ . This concludes the proof with

$$C_{16} := 2C^*. \quad (42)$$

□

*Proof of [Theorem 10.2](#).* Notice first that for all  $s > 0$ ,  $\mathcal{L}^s(\mathbb{R}^d) \supset \mathcal{C}^s(\mathbb{R}^d)$  so that we can rely on the proof of [4, Theorem 3.5]. This latter proof carries over word for word to the present setting after replacing  $\ell_2$ -balls by  $\ell_\infty$ -balls in the definition of the support of  $\mu$ . □

**Acknowledgement.** The author would like to thank Dominique Picard for many fruitful discussions and suggestions. He would also like to acknowledge interesting conversations with Gérard Kerkycharian. Finally, he would like to thank two anonymous referees and an associate editor whose constructive comments led to a full refactoring of the paper and largely contributed to improve it.

## References

- [1] Anestis Antoniadis and Dinh Tuan Pham. Wavelet regression for random or irregular design. *Comput. Stat. Data An.*, 28:353–369, 1998.
- [2] Anestis Antoniadis, Gérard Grégoire, and Pierre Vial. Random design wavelet curve smoothing. *Statistics & Probability Letters*, 35:225–232, 1997.
- [3] Jean-Yves Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. *Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6 and 7*, 2004.

- [4] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Stat.*, 35: 608–633, 2007.
- [5] Paolo Baldi, Gérard Kerkycharian, Domenico Marinucci, and Dominique Picard. Asymptotics for spherical needlets. *Ann. Stat.*, 37(3):1150–1171, 2009.
- [6] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146, 2002.
- [7] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3:203–268, 2001.
- [8] T. Tony Cai and Lawrence D. Brown. Wavelet shrinkage for nonequispaced samples. *Ann. Stat.*, 26(5): 1783–1799, 1998.
- [9] Christophe Chesneau. Wavelet block thresholding for samples with random design: a minimax approach under the  $L^p$  risk. *Electronic Journal of Statistics*, 1:331–346, 2007.
- [10] Albert Cohen. *Numerical analysis of wavelet methods*, volume 32 of *Studies in mathematics and its applications*. North-Holland, 2003.
- [11] Ingrid Daubechies. *Ten lectures on wavelets*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1992.
- [12] Véronique Delouille, Jürgen Franke, and Rainer von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *The Indian Journal of Statistics*, 63:328–366, 2001.
- [13] Bernard Delyon and Anatoli Juditsky. On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.*, 3: 215–228, 1996.
- [14] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 1993.
- [15] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, 1996.
- [16] David L. Donoho. De-Noising by Soft-Thresholding. *IEEE Trans. Inf. Theory*, 41(3):613–627, 1995.
- [17] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (3):425–455, 1994.
- [18] David L. Donoho, Iain M. Johnstone, Gérard Kerkycharian, and Dominique Picard. Wavelet shrinkage: Asymptotia? *J. Roy. Stat. Soc. B*, 57(2):301–369, 1995.
- [19] David L. Donoho, Iain M. Johnstone, Gérard Kerkycharian, and Dominique Picard. Density estimation by wavelet thresholding. *Ann. Stat.*, 24(2):508–539, 1996.
- [20] Stéphane Gaïffas. Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.*, 1:1–27, 2005.
- [21] Stéphane Gaïffas. Sharp estimation in sup norm with random design. *Statistics & Probability Letters*, 77: 782–794, 2007.
- [22] Stéphane Gaïffas. On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Statist.*, 11:344–364, 2007.
- [23] Włodzimierz Greblicki and Mirosław Pawlak. A classification procedure using the multiple fourier series. *Inf. Sci.*, 26:115–126, 1982.
- [24] Włodzimierz Greblicki and Mirosław Pawlak. Fourier and hermite series estimates of regression functions. *Ann. Inst. Statist. Math.*, 37:443–454, 1985.
- [25] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer, 2001.
- [26] Peter Hall and Berwin A. Turlach. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Stat.*, 25(5):1912–1925, 1997.
- [27] Wolfgang Härdle, Gérard Kerkycharian, Dominique Picard, and Alexandre B. Tsybakov. *Wavelets, approximation and statistical applications*. Springer Verlag, Berlin, 1997.
- [28] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, Berlin, 2001.
- [29] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.
- [30] Gérard Kerkycharian and Dominique Picard. Density estimation in Besov spaces. *Statistics & Probability Letters*, 13:15–24, 1992.

- [31] Gérard Kerkycharian and Dominique Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.
- [32] Michael Kohler. Nonlinear orthogonal series estimates for random design regression. *J. Stat. Plan. Infer.*, 115:491–520, 2003.
- [33] Michael Kohler. Multivariate orthogonal series estimates for random design regression. *J. Stat. Plan. Infer.*, 138:3217–3237, 2008.
- [34] Arne Kovac and Bernard W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statistical Assoc.*, 95:172–183, 2000.
- [35] Oleg V. Lepski, Enno Mammen, and Vladimir G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Stat.*, 25(3):392–947, 1997.
- [36] Gérard Malgouyres and Pierre-Gilles Lemarié-Rieusset. On the support of the scaling function in a multi-resolution analysis. *Comptes rendus de l'Académie des sciences*, 313(6):377–380, 1991.
- [37] Stéphane Mallat. Multiresolution approximations and wavelet orthonormal bases of  $L^2$ . *Trans. Amer. Math. Soc.*, 315(1):69–87, 1989.
- [38] Stéphane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic Press, 2008.
- [39] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Stat.*, 27(6):1808–1829, 1999.
- [40] James S. Marron. Optimal rates of convergence to the Bayes risk in nonparametric discrimination. *Ann. Stat.*, 11(4):1142–1155, 1983.
- [41] Yves Meyer. *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1992.
- [42] Jean-Baptiste Monnier. Non-parametric regression on the hyper-sphere with uniform design. *TEST*, 09 2010. doi: 10.1007/s11749-011-0233-7.
- [43] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9:141–142, 1964.
- [44] Michael H. Neumann and Vladimir G. Spokoiny. On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.*, 4(2):137–166, 1995.
- [45] Marianna Pensky and Brani Vidakovic. On non-equally spaced wavelet regression. *Ann. Inst. Statist. Math.*, 53(4):681–690, 2001.
- [46] Dominique Picard and Karine Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Stat.*, 28:298–335, 2000.
- [47] Sylvain Sardy, Donald B. Percival, Andrew G. Bruce, Hong-Ye Gao, and Werner Stuetzle. Wavelet shrinkage for unequally spaced data. *Statistics and Computing*, 9:65–75, 1999.
- [48] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Stat.*, 8(6):1348–1360, 1980.
- [49] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, 10(4):1040–1053, 1982.
- [50] Wim Sweldens. The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, 3:186–200, 1996.
- [51] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, 32(1):135–166, 2004.
- [52] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 1998.
- [53] Geoffrey S. Watson. Smooth regression analysis. *The Indian Journal of Statistics*, 26(4):359–372, 1964.
- [54] Yuhong Yang. Minimax nonparametric classification. I. rates of convergence. *IEEE Trans. Inf. Theory*, 45(7):2271–2284, 1999.
- [55] Shuanglin Zhang, Man-Yu Wong, and Zhongguo Zheng. Wavelet threshold estimation of a regression function with random design. *J. Multivariate Anal.*, 80:256–284, 2002.